# DOCUMENT RESUME

ED 144 268 EC 101 551

|  |  |
|---|---|
| AUTHOR | Mann, Lester, Ed.; Proger, Barton B., Ed. |
| TITLE | Research and Applied Theory in Special Education: Conversations with the Experts. |
| INSTITUTION | Montgomery County Intermediate Unit 23, Blue Bell, Pa. |
| PUB DATE | Aug 76 |
| NOTE | 335p. |
| EDRS PRICE | MF-$0.83 HC-$18.07 Plus Postage. |
| DESCRIPTORS | Conference Reports; Criterion Referenced Tests; *Diagnostic Teaching; *Educational Accountability; Educational Diagnosis; Elementary Secondary Education; *Evaluation Methods; *Handicapped Children; *Individualized Programs; Labeling (of Persons); Program Evaluation; *Research Design; *Research Methodology; Statistical Analysis; Student Evaluation; Teacher Evaluation |

ABSTRACT

The volume focuses on topics discussed at a symposium on research methodology in special education, with emphasis on the special educator's perspective. Following an introductory section (three chapters), are three sections with 17 chapters which cover the following topics: design of educational prescriptions--behavioral objectives and criterion expectancy levels; implementation and evaluation of educational prescriptions; diagnostic-prescriptive teaching and the Guilford structure-of-intellect model; diagnosis, labeling, and expectancy effects; evaluation of curriculum and instruction; accountability--measuring teacher effectiveness; considerations in standardized testing; innovations in testing--criterion-referenced measurement; reporting pupil progress; research utilization; complementary roles of descriptive and inferential statistics; implementing a research design with internal controls; matching and choice of appropriate control groups; reliability in special education measures, error of measurement, representative scores, and confidence intervals; measuring change; statistics--peculiar problems caused by the nature of special education research; covariance and regression to the mean; and research design and methodology--results of an older conference. A final chapter consists of a topical index with codes for such information as discussants' names and whether there are two or more sequential pages of related discussion. (SBH)

Documents acquired by ERIC include many informal unpublished materials not available from other sources. ERIC makes every effort to obtain the best copy available. Nevertheless, items of marginal reproducibility are often encountered and this affects the quality of the microfiche and hardcopy reproductions ERIC makes available via the ERIC Document Reproduction Service (EDRS). EDRS is not responsible for the quality of the original document. Reproductions supplied by EDRS are the best that can be made from the original.

ED144268

# RESEARCH AND APPLIED THEORY IN SPECIAL EDUCATION: CONVERSATIONS WITH THE EXPERTS

Completely Indexed

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Edited by

Lester Mann, Ph.D., and Barton B. Proger, Ed.D.

August, 1976

Montgomery County Intermediate Unit 23
Special Education Center
1605-B West Main Street
Norristown, Pa. 19401

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Barton B. Proger
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND THE ERIC SYSTEM CONTRACTORS."

Dennis U. Harken, Ph.D., Executive Director
Lester Mann, Ph.D., Director of Special Education
Barton B. Proger, Ed.D., Coordinator of Program Evaluation Services

EC 101 551

## TABLE OF CONTENTS

Preface....................iv
Sources of Funding....................v
Acknowledgements....................vi

## PART I

## INTRODUCTION TO PARTICIPANTS AND CONTENT OF BOOK

Chapter 1. How to Gain Maximum Utility from this Book....................2
Chapter 2. Current Interests of Participants....................7
Chapter 3. The Nature of Special Education Research....................10

## PART II

## APPLIED ASPECTS OF RESEARCH IN SPECIAL EDUCATION

Chapter 4. Designing Educational Prescriptions: Behavioral Objectives and Criterion Expectancy Levels....................19
Chapter 5. Implementation and Evaluation of Educational Prescriptions....................29
Chapter 6. Diagnostic-Prescriptive Teaching and the Guilford Structure-of-Intellect Model....................75
Chapter 7. Diagnosis, Labeling, and Expectancy Effects....................86
Chapter 8. Evaluation of Curriculum and Instruction....................100
Chapter 9. Accountability: Measuring Teacher Effectiveness....................117
Chapter 10. Considerations in Standardized Testing....................129
Chapter 11. Innovations in Testing: Criterion-Referenced Measurement....................158
Chapter 12. Reporting Pupil Progress....................173
Chapter 13. Research Utilization....................178

## PART III

## THEORETICAL ASPECTS OF RESEARCH IN SPECIAL EDUCATION

Chapter 14. The Complementary Roles of Descriptive and Inferential Statistics....................193
Chapter 15. Implementing a Research Design with Internal Controls Such as Adequate Sampling, Randomization, and Blocking....................207
Chapter 16. Matching and the Choice of Appropriate Control Groups....................222
Chapter 17. Reliability in Special Education Measures, Error of Measurement, Representative Scores, and Confidence Intervals....................240
Chapter 18. Measuring Change....................258

## TABLE OF CONTENTS (Cont'd.)

Chapter 19. Statistics: Peculiar Problems Caused by the Nature of Special Education Research....269
Chapter 20. Covariance and Regression to the Mean....282

PART IV

AIDS TO UNDERSTANDING

Chapter 21. Research Design and Methodology: Results of an Older Conference....299
Chapter 22. Topical Index....308

## PREFACE

This volume has attempted to delineate research methodology in special education as a peculiar area beset with many problems above and beyond the threats to classical designs used with normal children.

Above all else, this conference tried to keep things informal. There were no invited papers; we wanted only clear minds and spontaneous, fresh, thinking on a host of issues. We think these ends have been achieved.

Far too often regular education research experts fail to identify with what special educators feel are peculiar applied and theoretical problems in measurement and evaluation. We have long wanted to establish a better liaison between these camps, and we think the conference and the resultant proceedings represented here stimulate thinking along these lines.

The editors wish to acknowledge their appreciation to Victor Cogen, Ed.D., Bureau of Special Education, Pennsylvania Department of Education, who assisted us in the overall planning of the conference. Mann and Proger worked out a tentative list of topics to be discussed. Proger took the rough typewritten drafts of the transcribed tapes, separated them into content-consistent chapters, and indexed them for the proceedings. Mann assumed editing of the rearranged tapes.

Within these proceedings, the editors tried to preserve the original, natural manner of speaking of each participant to provide (hopefully) understanding and also the mannerisms of well-known figures in research, psychology, and exceptionality. Apart from this book's major purpose of providing a link between the research-methodological community (ipso facto, part of general education) and the special education realm, this document could also be considered semi-autobiographical; a bit of living history for future behavioral science historians!

The organization of the proceedings into more or less self-contained chapters is strictly arbitrary and was done at the discretion of one of the editors (BBP). The criteria for assigning one arbitrary label to the bulk of commentary in any one section was (a) that the label be descriptive of the majority of content in that untouched run of pages, and (b) that the label communicate with special educators. Others might prefer different breakdowns of the original, unedited, typewritten transcription, but we feel our present version is among the most readable possible.

L. Mann
B. Proger

Norristown, Pa.
August, 1976

## SOURCES OF FUNDING

These proceedings arose from a conference held February 9, 10, and 11, 1972, at the Marriott Motor Inn, City Line Avenue, Philadelphia, Pennsylvania. The conference was arranged and financed by Buttonwood Farms (3515 Woodhaven Road, Philadelphia, Pennsylvania 19154), by means of a federal grant to that school under Title I (Elementary and Secondary Education Act of 1965), P.L. 89-313. Specifically, Grant Number 48-7176-51-039 was entitled, "Prevocational Training for Emotionally Disturbed and Brain Injured Pre-Adolescents." The grant was funded during the period September 1, 1971, to June 30, 1972. Under that grant, as part of "Supplemental Professional Activities," the conference was entitled, "Pathology and Measurement Regarding the Exceptional Child."

Until this volume's publication in 1976, a great deal of typing has been donated by the Montgomery County Intermediate Unit in an effort to finally bring these proceedings to light. All of this editing that occurred in terms of content, style, grammar, and arrangement of topics, was completed by Mann and Proger over the course of four years.

## ACKNOWLEDGEMENTS

The conference has been the result of the coordinated efforts of many individuals. To the participants goes the greatest debt of gratitude; without their willingness to debate from all different angles the key issues in this conference, only a series of mundane, textbook-like thoughts would have arisen. Literally from early morning till evening during the three days, the participants freely exchanged the views that are here represented in the final draft of the proceedings.

To Mrs. Selma Carson, Director of Buttonwood Farms, goes sincere appreciation for the basic foundation planning of this grant: the initial arrangements and the funding! Mrs. Carson and her staff not only provided continuing support during the initial planning and actual implementation of the conference, but also later devoted a large amount of the Buttonwood secretarial time to the typing of some early drafts of the manuscript. Appreciation is expressed to Buttonwood Farms for providing the financial support of this effort.

Many thanks are extended to Dr. William F. Ohrtman, Chief, and Dr. Victor Cogen, former Advisor, of the Division of Special Education in the Pennsylvania State Department of Education. Without their encouragement and planning support, this conference never would have occurred.

To local Intermediate Unit staff, who had to suffer through several editions of the manuscript (the actual count is unknown at this point!), much gratitude is due. To Dr. Dennis Harken, Executive Director, is extended appreciation of his understanding and continuing support of local research efforts. During the early drafts of the proceedings, Mrs. Joan Shaffer valiantly struggled with making sense out of the original pre-transcription recordings of the conference. Later, Mrs. Frances McCann typed through some of the intermediate drafts. Dr. Robert L. Kalapos, Director of Regional Resources Center of Eastern Pennsylvania for Special Education, lent some valuable advice during the preparation of the manuscript. Finally, Mrs. Judy Rizzo and Mrs. Bernadette McMenamin were faced with final draft typing. Mrs. McMenamin was also faced with the gargantuan task of making sense out of all the various sets of index notes scattered throughout a few of the later editions of the manuscript.

# PART I

# INTRODUCTION TO PARTICIPANTS AND CONTENT OF BOOK

# CHAPTER I

# HOW TO GAIN MAXIMUM UTILITY FROM THIS BOOK

Overview: This chapter describes the overall organization of the book. The reader realizes how the chapters were formed and how to find a single specific topic not covered by an entire chapter.

## BACKGROUND OF THE MONTGOMERY COUNTY CONFERENCE

The Methodology Symposium was held at the Marriott Motor Inn in Philadelphia on February 9, 10, and 11, 1972. The purpose of the conference was to examine a prearranged list of research topics from the perspective of educators of handicapped children. It is safe to say that research methodology texts and applied statistics texts consider their respective domains only from the standpoint of mass education for the normal child. Clean research designs are the rule rather than the exception when dealing with normal children. However, with handicapped children, special educators find themselves in the opposite position of having rather "dirty" designs. Educators of the handicapped suddenly come upon peculiar problems in selection and size of samples, control of treatments, selection of appropriate statistical analyses, qualifications on results, and so on. Unfortunately, most statistics and design texts do not cover these "exceptions to the rules." It is precisely this situation which precipitated the Montgomery County (Pa.) Methodology Symposium. It is hoped that both researchers and practitioners concerned with handicapped children will come away with a better understanding of just what is "special" about special education research and how to handle these "deviations" from the mean designs and mean statistical procedures. With such a rationale in mind, the editors of this volume convened a group of research experts from the realms of both general and special education. Table I contains their names.

Heretofore, the only major conference to the editors' knowledge on special education research methodology was held in 1959 at the Woods School in Langhorne, Pennsylvania (see the September, 1959, issue of American Journal of Mental Deficiency). While the list of 1959 participants was indeed impressive, they were still nonetheless general education research people addressing themselves to the design and statistical problems of general education. Furthermore, the 1959 participants made their contributions by preparing beforehand a formal, "canned" paper or speech. Thus, the real flavor of special education research "messiness" never entered their thoughts. To avoid this pitfall, the present editors structured the three-day conference to occur spontaneously. What the reader is about to digest is the actual transcribed text, as it was spoken during the conference; only minor editing has occurred so as to remove profanity, thoughts that were cut off in midstream, repetitious dialogue, "ahs" and "ums," and the like. Thus, those present at the 1972 conference with special education background were able to put the general education research theorists on the grill with regard to sticky issues that are usually avoided.

## CONTENT OF THIS VOLUME

To gain a proper understanding of the editors' Gestalt for this volume, the reader should peruse the Table of Contents carefully. There are four main divisions to this work: (a) Part I: Introduction to Participants and Content of Book; (b) Part II: Applied Aspects of Research in Special Education; (c) Part III: Theoretical Aspects of

## TABLE I

## LIST OF PARTICIPANTS

| NAME | TITLE | AFFILIATION |
|---|---|---|
| Dr. Glenn H. Bracht | Assistant Professor | Division of Educational Psychology,<br>330 Burton Hall,<br>University of Minnesota,<br>Minneapolis, Minnesota 55455<br>(612) 376-7391 |
| Dr. Lawrence H. Cross | Assistant Professor of Education | Virginia Polytechnic Institute and State University,<br>Graduate School of Education,<br>Blacksburg, Virginia 24060<br>(703) 951-1641 |
| Dr. John F. Feldhusen | Professor & Chairman | Educational Psychology and Research Section,<br>Purdue University,<br>School of Humanities, Social Science and Education,<br>Department of Education,<br>Lafayette, Indiana 47907<br>(317) 749-2844 |
| Dr. Joseph L. French | Head, Department of Special Education | Department of Special Education,<br>Penn State University,<br>Rackley Building,<br>University Park, Pennsylvania 16802<br>(814) 865-6072 |
| Dr. Rosa Hagin | Research Associate Professor in Psychology | Department of Psychiatry,<br>New York University Medical Center,<br>560 First Avenue,<br>New York, New York 10016<br>(212) 679-3200, Ext. 4010 |
| Dr. Kenneth D. Hopkins | Professor of Education | University of Colorado,<br>Laboratory of Educational Research,<br>Boulder, Colorado 80302<br>(303) 443-2211 |

TABLE I

LIST OF PARTICIPANTS (Cont'd.)

| NAME | TITLE | AFFILIATION |
|---|---|---|
| Dr. Lester Mann | Director of Special Education | Montgomery County Intermediate Unit,<br>Colony Office Building,<br>6198 Butler Pike<br>Blue Bell, Pennsylvania 19422<br>(215) 643-7600 |
| Dr. Jason Millman | Professor | Educational Research Methodology,<br>Cornell University,<br>Stone Hall,<br>Ithica, New York 14850<br>(607) 256-3097 |
| Dr. Barton B. Proger | Coordinator of Program Evaluation Services | Montgomery County Intermediate Unit 23,<br>Special Education Center,<br>1605-B West Main Street,<br>Norristown, Pennsylvania 19401<br>(215) 539-8550 |
| Dr. Julian C. Stanley | Professor of Psychology | Department of Psychology,<br>Johns Hopkins University<br>Baltimore, Maryland 21218<br>(301) 366-3300 |

Research in Special Education; and (d) Part IV: Aids to Understanding.

Part I is necessary to set the stage for the main sections of this volume, Parts II and III. In Part I, for instance, the reader learns of the primary interests of the participants so that he can fully appreciate the perspectives from which the various recommendations are made. Needless to say, research, like teaching, can be considered somewhat of an art: there is no single "right" way to conduct a given experiment; different researchers have different recommendations depending upon what they place as priority. Also in Part I, a general discussion on what is "special" about special education research is given among the conference participants, although the full "answers" to this question can be gotten only by reading in detail Part II and III.

Part II holds wide appeal to special educators of every position because of its highly applied nature. No matter whether the reader is a teacher, administrator, psychologist, or whatever, this section will sharpen his insights on some of the pressing issues of the day. Diagnostic-prescriptive teaching, along with the related topics of behavioral objectives, educational prescription, and so on, are discussed in comprehensive detail. The perennial issue of diagnosis and resultant use or abuse of "labeling" receives close scrutiny. Accountability is afforded much attention in each of two chapters: (a) from the general standpoint of curriculum and instructional evaluation, and (b) from the more specific standpoint of teacher effectiveness. Finally, trends in both the new area of criterion-referenced measurement and the more established area of standardized testing are discussed.

For those readers who want to find out what special methodological modifications must be made to routine design and statistical philosophies already available from regular education practice, Part III presents valuable insights in many areas. It should be noted at the outset that if Part III were just a rehash of classical design and statistical theory, there would be no justification for it, in view of the plethora of texts available to the reader. Thus deliberate attempts were made every step along the way to bring out especially "sticky" design problems and then have the participants react to them. In this way, Part III should form a valuable adjunct to any existing text on design and/or applied statistics.

Finally, Part IV represents the editors' attempt to synthesize what is special about special education research. The highlights of each chapter are described and implications are noted. Also, the relationships between the 1972 Montgomery County Conference and the 1959 Woods School Conference are noted.

## ORGANIZATION SCHEME OF EACH CHAPTER

Each chapter has been structured essentially the same way so as to provide continuity and ease of usage throughout the volume. The first aspect of any chapter is the paragraph abstract (an Ausubelian "advance organizer," if you will!) that draws upon the major lines of thought within the chapter. Next, the naturally occurring dialogue of the participants is given. The reader should be aware of how this dialogue was generated. At the conference, professional tape recording and microphone equipment were used. Two recorders were always in operation to ensure complete coverage of what was said in case one machine would break down or would be "down" for a tape change. At the end of the conference, several secretaries were given an orientation to the voice characteristics of each participant. Several months later the transcription from tape to paper was completed. The editors then entered the arena and completely indexed each of the approximately 600 pages of double-spaced, rough-draft text. The indexing terms were geared at as specific a level as it was possible for the editors to think of. Thus, a single page might have five to ten terms attached to it. Once the detailed indexing was completed, it was possible for the editors then to see where a running series of perhaps thirty pages might deal with a single major topic (quite general indexing term) as opposed to the isolated occurrences of the highly specific types of indexing terms that were on each page. These major topics allowed chapters to be mapped out by the editors.

Finally, each chapter concludes by providing a list of references. This list was selected by the editors as being representative of the field of endeavor that the chapter covers. The list of references is not to be considered a comprehensive one but hopefully a representative one.

## FLEXIBILITY OF USAGE OF THIS VOLUME

The editors feel this volume is suitable for several types of users, among them: (a) administrators and teachers in public and private school systems who wish to see how research of an applied nature can be carried in their local settings to aid decision-making; (b) researchers and psychologists who wish to hear innovative solutions to "sticky" design problems not commonly discussed in existing texts; and (c) as a supplementary book of readings/conversations for any graduate student in special education.

Apart from the specific audiences to which this volume appeals, the editors see a common mode of utilization. First, the reader should of course scan the Table of Contents to find the general area of interest and to make sure he is not missing any related chapters. Second, the reader should digest the introductory paragraph abstract before each chapter to make sure the chapter is what he thinks it is. Third, the reader proceeds through the actual conversational dialogue. Fourth, the reader might wish to consult one or more of the references for additional reading listed at the close of each chapter. Fifth, the reader should turn to the highly specific index and search out all those subtopics that pertain to his general area of interest; in this way, the reader will be led into other chapters that deal with his topic as a secondary (but still highly relevant) issue. The editors feel the detailed subject index, which may appear labored at points, is a unique feature of this volume which, by its very nature of informality, needs a great deal of structure present to retrieve topical information.

# CHAPTER 2

## CURRENT INTERESTS OF PARTICIPANTS

Overview: The reader hears from around the conference table what each participant is presently concentrating upon. This information is designed to provide the reader with the perspective from which each participant is making his subjective recommendations. Hopefully, such information will allow the reader to attach differential weights to what each speaker is saying, depending upon how closely the reader feels he is willing to identify with the speaker.

Mann: We are ready to start. Dr. Hagin has suggested that we pass briefly around the table and simply identify current research activities. This is to orient our audience. May we start with you, Glenn, and proceed around the table.

Bracht: There are two things that I am presently working on. One is procedures to use in observation. How do we observe the educational programs that we are carrying on! This requires, you know, either an external observer or someone who is a participant in the activity making observations, let's say, at the end of an activity either the student or the teacher doing it, and it includes the different things that are observed. We are looking at the cognitive components of the instructional setting, at the affective components and at the management of the components. The other thing that I am presently working on is an evaluation of the properties of training programs for science and/or mathematics teachers—institute research sponsored by the National Science Foundation and we are just presently starting an evaluation of that and of the effects of the programs.

Millman: The two activities that are taking the most of my time are (1) developing programs to train people to appraise research critically and (2) looking at some of the problems in the measurement of teaching effectiveness.

Hagin: We're part of the Department of Psychiatry—a multi-disciplinary unit including teachers, psychiatrists, social workers, and psychologists. Our work goes on primarily outside the hospital walls in the schools of lower Manhattan. We have two intensive projects, one in lower Manhattan where we are working with an intact group of children to understand their problems better and possibly to intervene with those children who do seem to be learning. We are also learning about pre-schoolers who are potential learning disabilities. I am not sure this has reached the point of research, but we are exploring what can be done to intervene early with them.

Mann: I am going to return to Jay for a moment, for him to discuss his mini-teaching project.

Millman: One approach to measuring teaching effectiveness is to assess the attitudes and abilities of the teachers involved and in order to develop measurements which really permit comparisons among teachers, we ask them to teach a common set of objectives. These objectives generally deal with original types of materials in order to minimize the previous content knowledge that the teachers may have. We call these instructional assignments mini-lessons and we are using them in this evaluation. We are also preparing, if you will, kits of these mini-lessons and suggestions how a teacher, either in isolation or with others, in either preservice or inservice education, may use to improve her methods of teaching.

Mann: Ken?

Hopkins: One topic of immediate concern which I am giving some attention to is the effect of the reliability of the dependent variable on the resulting analysis. I am also doing a robustness study dealing with the consequences of violating a particular assumption, still another study pertains to consequences of various items selection procedures and with the effects of various types of item selection strategies on the resulting qualities of the measure.

Mann: I think Bart and I here are working essentially on a criterion-referenced in mastery learning approach to reading and mathematics. This is being carried out with learning disabled children. Bart is also involved in an evaluation effort for the Commonwealth's special education system and I direct the Special Education Program for Montgomery County (Pennsylvania) as my main pursuit.

French: Well I am out of it, that is, during the last couple of years I have been trying to play the roll of an administrative clerk in getting some of our graduate programs reorganized at Penn State. Prior to that I had a study of high school dropouts with IQ's of 110 and above which was a descriptive study. Another one of my interests has been in assessment problems of the cerebrally physically handicapped and in trying to work with their learning problems. We are also trying to organize a center for educational diagnosis and remediation that will get at diagnostic and prescription problems.

Feldhusen: No. 1 study that I have been involved in for a long time... the children who are identified as persistantly highly aggressive and disruptive in the classroom and I am still working with that, and No. 2, a study at present of several modes of testing for problem-solving abilities among disadvantaged children along a continuum for highly abstract form, to quite concrete 3D form. Those are the two things at present.

Dr. Julian Stanley joined the conference at a later time. His interests at present are founded upon problems of identifying and training gifted and talented secondary students.

## CHAPTER 3

# THE NATURE OF SPECIAL EDUCATION RESEARCH

Overview: The question implied by the title of this chapter is a huge one. Indeed, the entire book addresses itself to this issue. Thus, the purpose of the present chapter is only to present some top-of-the-head reactions from the participants in this conference.

Mann: He enjoyed the last one. This one may not be as much fun... might be more intelligent conversation though.

Sinkler: It was fun the last time. O.K., now what the story is here. We are recording everything and as you were saying, it will be edited down and you can work that out—those things that are to remain in or not, but the point is we want to make sure that every single word you say is very intelligible because they are going to transcribe from these tapes, so try not to talk over other people. If someone else is talking and you want to make a point, wave your hands or something that doesn't make too much noise on the tape and it will be a lot easier to transcribe. As far as the microphones are concerned, just forget they are there. They are good mikes and they will pick you up. Just completely forget that they are there. Speak in a good, strong, tenor voice. Aside from that, don't worry about anything as far as the microphones are concerned. Don't worry about what direction you are facing. You can sit back if you want to. I can take care of all that. There is no problem there. We are going to have both recorders going the whole time which means in case anything goes wrong with one of them, which is completely unlikely, we still have the second tape. Every precaution has been taken so that really nothing should or can go wrong. I'd like to go around the table now very quickly and have each of you identify yourselves so that the people transcribing can get the voice and the name connected and also I will be able to get a quick idea of how loud you are talking. So if you will begin on the left...

Sinkler: I'm happy. Dr. Mann, it is all yours.

Mann: We hope that the next several days will be productive, in terms of answers, to a variety of issues. Special education in general, especially childhood exceptionality, has lagged behind in terms of research sophistication in comparison to regular education. We hope this conference will assist in bridging that gap. I would hope for a free-flowing discussion. I don't think we have to stick to any general topic. We will splint and paste later on, but let's begin with the general issues of work with handicapped youngsters; extend this to culturally deprived and finish up in the final day with the gifted, if possible. Bart will express some of our concerns that you, in general education, perhaps are not fully turned into. Among other things, we are concerned with problems in special education curricular package research. We are always talking about building prescriptions for kids—individualized instructions—tailoring teaching to their individual learning styles and individual differences, and the conformance with every last feature the test provides, as we come up within a diagnostic evaluation, and the question that always comes up again and again is how effective are these specific little training approaches in the prescriptions that you put together for kids? Welch and others have written reviews of the problems in such evaluations. Bart, can we pull back a little bit to a general overview first, of the field? Can we ask first, if there is a specific area of research we could deliberate for the exceptional child? Is it legitimate to conceptualize a separate field of research for special education, or for exceptional children?

Feldhusen: Your question, in part, sounds as though you are asking, is it legitimate to attempt research at all in this domain, and yet, I don't think you are asking that. I'd like to pass, because I am not certain, at this point, that I understand.

Hopkins: I'll take your question, Les. Of course, in the research endeavor, one tries to find out information to which he can generalize at some particular population. Perhaps the question is, are people in special education interested in having the kind of information that they can use on the population they deal with. I think, in many areas, the real decision makers are really unconcerned about such information and they don't use it for educational improvement. Not only at the academic institution, but also in practice, the decision makers are relatively uninfluenced and unconcerned about research evidence, so while in theory research should improve practice, but it rarely does. I think, in the area of special education, the basic problems are not unique in terms of research strategy. There are, of course, unique definitional and measurement problems, but I don't see anything fundamentally different in research for general education.

French: I would agree with you, Ken, that the processes shouldn't be any different, but the problems are. I am a little concerned about the generalizability of results in work with the handicapped, because the characteristics of the people we work with, in many instances, are so different from others that it is really hard to generalize, in many instances. As Bart was talking, I was caught up with the idea of curriculum development and reminded of the kind of validity that you don't think of very often, but which one guy referred to as "cash validity." A curriculum, or material, is valid if it sells well, and in special education, we are plagued by a lot of materials that do sell well, and I think it is up to a number of researchers to determine exactly how well these materials really work with kids. They come on because there are legitimate needs...something that will help the kid who has a problem. So then it gets into the business end, and is general and used indiscriminately. We have to validate the approaches.

Proger: We are, particularly in special education, plagued with a vast array of instructional materials and programs, and there is a real problem in trying to demonstrate what does and doesn't work. Unlike general education, where you typically evaluate a whole year's work, or activities, and thus have an extended study of instructional components along the way, most of the materials used in special education are for much shorter durations. Now, how can we evaluate the effectiveness of a program that lasts but one or two weeks?... relative to any other program. So, though I agree that there are generalizable approaches, like Ken said, in any type of research endeavor and in methodology in general, I think there are also some peculiar hangups in special education materials validation.

Feldhusen: It seems that our discussion is relating to the battle between conceptions of research, as opposed to conceptions of curricular evaluation or new approaches to evaluation, and I am sure that everyone around the table has read the PDK volume on evaluation that Dan Stufflebeam edited, where the argument, of course, is very strong that these are separate domains. This is a reaction to Bart, here, that we ought not to run these two domains together—the domain of curriculum development and evaluation, and the domain of research. Now, others may disagree around the table, but at least we ought to begin by recognizing that there is a substantial scholarly issue, if you like, at the present, with certainly Mr. Stufflebeam arguing that there are quite different assumptions guiding the methodologies involved in evaluation, and in instructional development. I have said my say.

Bracht: I think it would be valuable for us to get on some common ground. Several references have been to evaluation of curricular materials. Is that what we mean by research in special education, or are we talking about other things?

Hopkins: I would like to respond, in a sense, to applied evaluation and research. I don't happen to believe that they are basically different, but I do think they are different points on a continuum. However, when we talk about curriculum evaluation...we're talking about evaluation of something that has a life of only a few years, and I think that often, in terms of cost and benefit considerations, it is very inappropriate to think about using classical experimental paradigms for getting at such kinds of questions when the materials are going to be replaced by something else and in not that distant a future.

Mann: I agree that if you want to go to curriculum evaluation, you are never completed—never off the treadmill. If we keep evaluating products, we never get to the job of teaching the children.

Feldhusen: Well, I don't think that is so. Certainly, well-developed models for curricular evaluation exist, like that which

Richard Schutz has developed for reading programs from the regional lab...as a result, the program has undergone a completely different kind of development from the typical commercial thing that has been developed in the past. There is another project, too, where Jack Crawford has tried to trace the complete continuum of basic research that undergirds that project which, by stages, has finally led to this viable and now seemingly well validated reading instruction program. I think such approaches create a quite different situation from what you are talking about, Les and Ken ... namely commercially developed, essentially unvalidated instructional packages that come and go and proliferate upon the market. I endorse the types of models that Schutz and Baker spell out in their new book which provides comprehensive and valuable processes for curriculum materials development that can result in different and new and well validated kinds of instructional material. So I would have to disagree with the ephemeral come-and-go assertion that has been made here. I am hopeful that what they are doing will lead to something completely different in the way of instruction materials.

Mann: But your comment still doesn't negate the fact that it is really a very difficult task even to keep up with much less validate, new curricular approaches.

French: I don't know if this is really the topic that we are here for but it seems that the cost of curriculum development and evaluation is exorbitant when you do it correctly—and I don't know if society is ready to pay the price for it, and when you come to special education, there isn't much of a market for the materials. Now if you've got something that applies to millions of kids, then it is easy to write off the enormous development cost for a program but when you have something that applies to only hundreds of kids, well then the cost per copy is enormous and there isn't the development money around for such things and so I think for a long time there will be people who will try to cut corners tremendously to try to get out materials that at least seem to meet such needs.

Proger: There is an interesting parallel here, I think, between the lack of proper validation of materials and the lack of adequate validation of newly released tests. You have only to pick up the manual of even the most commonly used tests today and you will find the validation is bad. The standardization of the tests may be great, you have a tremendous sample perhaps very well selected to represent all parts of the country and all types of characteristics but the validation is often very, very inadequate and it is only after many years on the market that you can say we are beginning to get some evidence. For example the ITPA or the WISC after many, many years begin to build up an accumulation of information but when they are first released, validation is pretty poor.

French: I think that the importance of curricular, technical, program evaluation and validation is critical of certain packages in special education to determine how effective they are. It is very difficult to engage in other...in more formal research activities if you don't know how good the things are that you are working with. The assessment devices to determine whether or not the kids have a visual perceptual problem. If they are unreliable or invalid and you have sorted your groups on that characteristic, we are in real trouble with our research. If the materials that we work with to remediate problems aren't validated, how do we know whether it was the materials or the characteristics of subjects that create our results?

Mann: You make a very good point. I was just reading about exceptionalities in Eastern Europe of special education and one of the real problems there is that they have a heck of a time comparing groups because there is very little in the way of standardized tests. It is against the Soviet philosophical grain to use standardized achievement or ability tests. So when their people talk about certain kinds of children's characteristics, they are not portable or translatable from place to place.

Hagin: The... definitions of the populations, of course, is critical. In the area of reading, for instance, studies report on children who were in the lowest 10% in population and they find things very different from what we might find where we use our type of expectancies as a determinate of reading disability. So they may be talking about children who are slow learners because of poor reading skills and we may be talking about children who are slow learners because they have language disabilities. So the findings are very different from both areas and I feel that somehow if research is to have an impact on what we choose to teach children, we have to define the children who have been studied in research more precisely and using maybe not just tests in defining the characteristics. Standardized achievement tests that are inappropriately standardized in terms of the large city populations are still used within...we keep using the same tests and complaining about them.

Feldhusen: The implication of what you are saying is that there are special characteristics of sub-groups of children that interact with the nature of materials or programs...not that these latter lead to differential performance on the part of that group of children or some group as opposed to some other group. I would certainly think within the special education domain that if we could not find such interactions anywhere else, we ought to find them here and yet this area of investigation has been very discouraging. This is trying to find if there is special material that works best for this group or type of youngster that works badly for another group is there, in other words, what we now know as ATI or Aptitude Treatment Interaction research, and it is very discouraging. I have to be very hesitant because Mr. Bracht, here, is the author of a review in this domain recently. His review, along with that of the Cronbach-Snow review, suggests that as nice as it sounds, the whole world of research is hard pressed to come up with demonstrations of differential effectiveness, particularly of instructional materials, for subgroups of youngsters, so while the whole world speaks of individualizing instruction one has to be very discouraged about our ability for the research world to produce such evidence. Now I have made an assertion here...others may disagree with me.

## REFERENCES

Saloman, G. Heuristic models for the generation of aptitude-treatment interaction hypothesis. Review of Educational Research, 1972, 42, 327-343.

Cronbach, L. J., & Snow, R.E. Individual differences in learning ability as a function of instructional variables. Stanford University, 1969.

Bracht, G. H. Experimental factors related to aptitude-treatment interactions. Review of Educational Research, 1970, 40, 627-647.

# PART II

## APPLIED ASPECTS OF RESEARCH IN SPECIAL EDUCATION

## CHAPTER 4

# DESIGNING EDUCATIONAL PRESCRIPTIONS: BEHAVIORAL OBJECTIVES AND CRITERION EXPECTANCY LEVELS

Overview: Diagnostic-prescriptive teaching has received a great deal of attention in special education. Several chapters in Part II of this volume deal with this topic and Chapter 4 concerns itself specifically with the derivation of educational prescriptions. For such a task, one needs a psychoeducational diagnosis. One needs also to establish the behavioral objectives that constitute the goals of the "prescription." Chapter 4 also pays attention to the sequencing of instruction and thus enters the realm of task analysis.

Hagin: This is an old argument we have engaged in the *Journal of Special Education*.

Mann: But an important one. Let's give a concrete example. You have a child who has a reading problem and has been given certain tests which found him to be perceptually deficient. He would go into a total program of perceptual training, if I interpret Rosa's position and that of the other process people correctly.

Millman: I had a kid like that. He was given these tiles to work with. Is that the debate between Rosa and you and another group?

Mann: Well, it is a debate between Rosa and me, but one debated by the entire field. Can we genuinely teach reading by approaches that are not specifically reading diverted.

Hagin: What would you do, Lester? You are against having the child work a walking board and you would not have him assemble tiles, and yet he has not learned by the conventional reading programs in the classroom.

Mann: I would do a task analysis of his problems and the reading sequence he is confronted with. It may well be that learning to assemble tiles is somehow propaedeutic to learning how to read later on or whatever, and if so, I will try it out... but I don't have to make any assumptions about the child's perceptual state in so doing; I simply see if this activity is relevant to his training through testing it out. Does it work or not? I don't presume I am teaching a basic skill or ability or whatever, all I want to know is: does it teach him to later on read correctly or do whatever.

Hagin: How would you do this?

Mann: By establishing the child's initial competancy level in reading or writing or whatever and proceeding upward from there. We are in fact doing this, using a traditional curricular and have been successful.

Proger: This is as opposed to working with perception, etc...in specific language skills. We find that a curricular focus does it for learning disabled children at least.

Mann: Should there be guidelines established for the termination of formal instruction for the handicapped? Let me give you a concrete example. You have an educable, retarded class and are able to teach the children to a certain level, using such and such techniques. The child reaches fourth grade reading level. Is it at this point, legitimate to push for a higher level? Do we stop here? Is this a research problem at all, or a philosophical issue?

Hopkins: It would seem that using past performance as a baseline for evaluating present performance would create the most meaningful criteria for developing expectancy levels. Absolute standards that we could go by would be very desirable, but are very difficult to operationalize, and in lieu of them, using past performances as a baseline for expectations is, I think, a very useful strategy until we get to the point that we have absolute standards. Until that time, I don't see it being ever appropriate to stop trying to improve.

Mann: John, can we establish output standards? Is this something to be worked for?

Feldhusen: I think, in truth, this can be very much an empirically determined proposition. If the situation were reading, for example, it seems eminently possible that one could design a study to determine the necessary levels of reading proficiency that a special child would have to attain if he was to...if your goal was to try to get him prepared to go back to a regular classroom or a special job training and function adequately...the minimal level of reading ability he must have, to learn adequately and survive emotionally and socially, etc. I see such things as empirically determinable things.

Mann: Jay?

Millman: Yes, I think there are a number of ways in which you could establish expectancy levels and I think John hit on one of them and, in my opinion, the one that is the most satisfactory one if we could get the data. Mainly, you determine the standard by seeing what performance is in order, to do the next step, whatever that might be. Lacking such data, there are other ways of proceeding. If you are working with a criterion-referenced situation in which you think you can define a universe of items, you can conceptualize the possible items and then determine how many of them you would need to ask on a test in order to get certain kinds of answers about levels reached...I guess you could work backwards and say that you could figure out, for example, what performance level you would require in order not to make certain kinds of errors of misclassification. By the way, you could look at the kinds of classification and ask yourself the question: Which type of classification is the worst to make? If you are afraid of saying that somebody really knows his stuff when he really doesn't, that really bothers you because it might mean putting him in a situation in which he is not prepared. Then it seems to me you would set a higher standard and expect more of him than if you were in a situation where having him referred back for remedial stuff doesn't present any great problems. In that case, you would be able to tolerate, it seems to me, a lower standard. If you are not sure here, you could move to—just sort of do a logical analysis, I guess. That is what they recommended in the past. If you go back and read Daniels, for instance, that one question covering measurements...

Mann: Logical criteria?

Millman: Well, you ask yourself about what are the kind of things that a person who is going to be reading—how many of these items does he really need to know—something like that. There is an extension of this which the people in the medical profession have used, primarily the University of Illinois; they have used what they call the minimum-task level technique. They have some interesting ideas about pass-fail. A lot of people are worried about what should be the passing score and their approach, which is, you look at each item and determine the probability that when somebody is going to fail —what are the possibilities that they are going to do something that would say, kill a patient, whatever the question is, and they look at those probabilities and on the basis of those probabilities, they wind up with minimal task levels. Now that could lead to some very unfortunate kind of things. For instance the person, who on a two-item test might know one of the answers correctly or an objective on one of the answers but also not "know' something that results in his killing a patient, and he would pass, whereas somebody else would give exceptions to the questions that would be legitimate but not the "correct" answer for either of the two items and he would fail. So, there are some flaws in that particular technique, but it is an example of a procedure in which you look almost item-by-item and you ask yourself: What do I require a person to know—to be able to do this, before I pass him?

French: Is the problem with the technique, or the items that are employed within the technique? As you were saying, it doesn't work.

Millman: Well, it doesn't work because they make a distinction between three kinds of answers. Answers that are correct—answers that are incorrect, but, you know aren't too dangerous to be marked that bad, and then items that are completely unacceptable kinds of answers. They only give point if you got correct answers, but they base the passing score on the ratio of these different kinds of errors. While the idea is good, it doesn't quite count out mathematically. But I'm sympathetic with the approach because it was—it tried to put a cutting score on a criterion-referenced observation. I've written this. I haven't look at that paper and probably won't until two weeks before AERA. I tried to summarize the literature on the various approaches to setting cutting scores on criterion-referenced measurement, and this was an example of one general point. But, to summarize, I think probably the most, to me, defensive way in which you determine what should be expected of a child before you move him ahead ought to be based on research, if shown what kinds of performances are needed in order to be successful later on, and if research shows that a student can miss most of these concepts but still deal adequately with this next step, then it would seem to be you should have a lower requirement than one which says if the guy is all messed up on these concepts he is really going to be messed up when he gets to this next thing; then you should have a high cutting score or a high testing score on the first hand.

Proger: The type of research you just suggest here suggests genuine knowledge of the infinite number of steps and combinations of previous knowledge regarding the sequence of activities for courses of training and study. Do you think we will ever get to the stage where we can say...Yes, we have gotten such a huge body of data together that we can begin to make such logical or empirically-derived conclusions as to what is essential—let's say for an EMR kid to know before he goes on to a higher level of basic reading skill—phonics analysis.

Millman: I think if I wanted to train a person how to do long division —in arithmetic—I would know, and kind of quick, if he were ready—say multiplication or subtraction. I think there would be a chance—a big chance—that if he can't pass that test then he is going to have trouble with long division.

Proger: I think maybe I'm misreading you. You are talking about giving tests to say these kids passed or didn't pass and then maybe do some research as to whether the kids that failed would still be able to handle the next stage in instruction adequately anyway. I'm thinking more of a factorial analysis or a factor analysis of different, essential, minute skills that go into the performance.

Millman: That's messy, too, because even if you find the relationship between a score on some sort of a possible prerequisite test and the later performance that you are really interested in ...it doesn't necessarily mean that if you—if the child is pushed up to the criterion level that then he will be able to handle the later test.

Hagin: We are talking about minimum levels, minimum—yes? Some of those things I am interested in having children reveal that they have fully mastered and retained the lower achievements before they proceed.

Mann: That is a good point. Rosa is saying that learning-impaired may disrupt kids if they don't have continuing practice with a particular skill; they will wash out. So, perhaps we need more absolute mastery of certain basic skill components.

Hagin: The children themselves have told us in follow-ups. "I was just beginning to understand it when the teacher went on to something else." They need more, perhaps, than the ordinary mastery that the kid going through a regular curriculum might need.

**Feldhusen:** Well certainly, the most valuable model now for the kind of thing you are talking about, I think, is the Gagne model and I don't think that it is appropriate to apply any conception of minimal attainment or performance within that model. I think the attainments implied for each task performance level within a Gagne type hierarchy quite to the contrary, implies mastery within each of his blocks. My recollection is that something more like a very effective performance is implied. The second happy circumstance of the Gagne type hierarchy is that one always sees the higher levels in the hierarchy as representing more complex higher level cognitive functions (ultimately problem-solving mastery) at the top of the previous sequence, so that I think that if one accepted that model for what we are talking about here—and certainly a tremendous amount of research has been generated by it, not all of it supportive, then something different than minimum performance at each level is implied, and remember that Gagne's model is not a linear step-by-step upward model either, but has many different components coming in, in a sideways fashion, as one works up to higher level skills. So I think that as a research model, the Gagne hierarchy may have application within the special education domain.

**Bracht:** The concept of mastery, I think, is a very important one. When we are referring to very specific skills—you have it or you don't have it. For example, the skill of adding two one-digit numbers—adding 1 + 6, 7 + 3, etc., that is, a skill which a student probably has or doesn't have and there are a number of tasks (that is, a number of problems of two one-digit numbers)—a limited number of those tasks from which we could select tasks to test the student. And I think we could say, from the student's performance on these tasks, that he has mastered this skill or he has not mastered the skill. When we move to broader outcomes, I think it becomes more difficult to say a person has mastered something. Perhaps the domain we have in mind is elementary school mathematics. I don't think we could ever develop a test that would measure whether a student has mastered "it" because "it" consists of so many things that a test could never fully convey, would be almost a curriculum and so when we move to the measurement of broader things I don't think we can talk about mastery...only for the very specific skills, and I think that there is a very real need for measuring such skills and that is where we can use criterion-referenced measurement for those very specific skills.

**Millman:** I would like to cut in on both John and Glenn's comments. I think we have to be cautious in talking about "mastery."

I remember Lee Cronbach making the point about mastery not really being a correct word as we tend to use it. You know, the notion that you have either mastered, or not mastered, something. Even with the two-digit numbers thing, I think of all various ways you could see if kids could answer such problems: different numbers, horizontally vs. vertically, word problems! I think, except with perhaps the most trivial kind of thing, I'd say that probably complete mastery isn't really the notion we had in mind. Now, in reference to the Gagne thing. I believe that with the Gagne thing, it is a lot more horizontal and a lost less vertical than he, or we, are probably willing to admit. I suspect that there is not many things that are absolute prerequisites for other things that are to be taught and learned. You need to know; supposedly, you should know, how to solve a simple linear equation before you solve a problematic equation. Yet, I think I could teach a person the general procedure for solving quadratics and then have him learn how...to solve a linear. I think it isn't necessarily so that you must learn this, before you learn that. In fact, much of the hierarchical nature that we build into our discipline just represents a way of training a teacher, perhaps, rather than a necessary condition. This being the case, I would logically argue that maybe one should not require a high level of performance on a lot of different things, if one is willing to try on for size the higher instructional goals and sequences, rather than keep the child from ever reaching that level.

Mann: I think this is important to special education and for the developmental hypothesis upon which much of special education is...You work or train to accomplish certain development levels before going to the next—work with the talking typewriter shows, if nothing else—that you can teach a kid at least at the level of word recall, before perceptual skills have reached maturity and recent work reported with Thalidomide children indicate a number who, despite impaired or limited spatial and movement experiences, have progressed adequately, in terms of learning mathematics skills, and so on.

Hagin: I would like to ask Jason why he was teaching this child quadratic equations. What I am asking is...Is your goal somewhat more limited than it should be, if you are only going to teach towards this specific goal?

Millman: I am not sure I follow.

Hagin: Remember, you said you would teach—you said you could teach quadratics, even if the child didn't know any equations—wouldn't that be a narrowed and limited goal?

Mann: But if he can do quadratics, couldn't he also do linears? I would think so. If, instead of teaching, say the alphabet, if you teach the kid the word, wouldn't he be able to discriminate the alphabet letters later on, too? In other words, if lower level skills teach you higher level skills later on, why can't you go conversely? Teach him the higher level one which encompasses the lower level skills...

Millman: The point I was trying to make, Rosa, I think—actually there were two points. The first being that I suspect that there is a strong belief in the hierarchical structure of knowledge, i.e., that certain things have to be learned before certain other things; like you have to crawl before you can walk. But much of this may not actually be the case. In fact, one reason that we do consistencies, in terms of developmental stages, might be because of the way that people have taught and trained children—the stages may be partial artifacts of the system.

Feldhusen: I have to agree with what Jay has said, but I want to make clear in respect to hierarchical analysis, that it is not simply a faith proposition. It may begin with faith, but aside from the theoretical assertions of Mr. Gagne, it has also been empirically validated. Now the practitioner may take the concept of learning hierarchies perhaps too seriously and come to erroneously perceive fixed and inescapable structures and sequences where they are not. The researcher, however, begins by trying first logically and subjectively to analyze this chunk of function, discipline or whatever it is, into a set of ordered tasks and then always submits them to empirical tests to determine if there are indeed such hierarchies.

Millman: But how do you test it? If you teach everyone the same way you get a particular hierarchy that may be a function of how you taught them...rather than a manifestation of some inherent hierarchy. I've looked at all this research on the taxonomy, i.e., and don't find them at all convincing.

Feldhusen: You mean their hierarchical nature is not convincing?

Millman: Not at all...I would be very much surprised if, for example, in Gagne, he has discrimination and concepts and classifications, whatever, and I'd be very surprised if you'd see this all emerge in learning, except through the researchers' contrivance.

Feldhusen: Well, first of all, I agree that the attempts to validate hierarchical structure within the cognitive taxonomy have been quite unsuccessful. And by no means have the many efforts to validate the structures suggested by the Gagne approach been successful. There is much imperfection in what seemed to Gagne, and his original approach toward a logical structure, Paradise, and so hierarchies do not come out perfect—no! And we can go over to the programmed learning domain and pick up a total contradiction to substantiate. That is, those several studies which seem to suggest that you could just throw learning material into complete random disorder and accomplish proper learning. Think, for example, of one study where Skinner's book, The Analysis of Behavior—a programmed text—was gone through, not strip-by-strip the way you are supposed to, but in grossly scrambled fashion, and yet still seemed to result in learning the Skinner method very well. So, there are illustrations obviously at the gross opposite end of the spectrum where presenting instruction in totally randomized order still seems to produce learning, but I think that for the kind of tasks that are described in the Gagne-type hierarchy, there is some support. Mathematics and the sciences—they are the best support for the presence, of some degree, of hierarchical nature to the learning sequence. There is less evidence in other curriculum areas.

Millman: Well, that can be a matter of artificial hierarchy which would probably be most justified, and therefore have conformance also—the material you teach fits a presumptive hierarchy.

Mann: In reading, it is extremely difficult to scientifically, or otherwise, structure it legitimately, once past the basic word-attack skills...the high order components in reading are difficult to even clearly establish, much less arrange hierarchically, and many people read all sorts of different ways and equally effectively.

Feldhusen: A very good thing really happens at the first stage in the Gagne analysis, in any event, and that type of analysis, as opposed to the Bloom type, is that the first stage is always a concerted effort to think through logically or rationally, as Jay suggests, the possible relations. It is first, really an effort to define manageable separate tasks; then, secondly, to try to conceive of a hierarchical arrangement of those tasks and the component learnings that are involved. That as a prelude to designing instructional material or instructional procedures, might indeed be quite a valuable contribution in itself.

## REFERENCES

Duchastel, P. O., & Merrill, P. F. The effects of behavioral objectives on learning: A Review of empirical studies. Review of Educational Research, 1973, 43, 53-69.

Gagne, R. M. Learning and instructional sequence. In. F. N. Kerlinger (Ed.) Review of Research in Education: 1. Itasca, Ill.: F. E. Peacock, 1973. pp. 3-33.

Hastings, G. R. Independent learning based on behavioral objectives. Journal of Educational Research, 1972, 65, 411-416.

Heimer, R. T. Conditions of learning in mathematics: sequence theory development. Review of Educational Research, 1969, 39, 493-508.

Jenkins, J. R., & Deno, S. L. Influence of knowledge and type of objectives. Journal of Educational Psychology, 1971, 62, 67-70.

Kibler, R. T., et al. (Eds.). Behavioral objectives and instruction. Boston: Allyn and Bacon, 1970.

MacDonald, J., & Clark, D. F. Critical value questions and the analysis of objectives and curricula. In R. M. W. Travers (Ed.) Second handbook of research on teaching. Chicago: Rand McNally, 1973, pp. 405-412.

McNeil, J. D. Forces influencing curriculum. Review of Educational Research, 1969, 39, 293-318.

Popham, W. T. Objectives-based management strategies for large educational systems. Journal of Educational Research, 1972, 66, 4-9.

Quilling, M., & Otto, W. Evaluation of an objective based curriculum in reading. Journal of Educational Research, 1971, 65, 15-18.

White, R. T. Research into learning hierarchies. Review of Educational Research, 1973, 43, 361-375.

## CHAPTER 5

## IMPLEMENTATION AND EVALUATION
## OF EDUCATIONAL PRESCRIPTIONS

Overview: In Chapter 5, one turns to another aspect of educational prescriptions: their implementation and evaluation. In particular, the discussion is couched in terms of aptitude-treatment interations (ATI) philosophy; a special category of individualized instruction that has particular relevance for special educators.

This chapter discusses the problems and heretofore poor showing of ATI research. One issue that is raised is the specificity of the diagnostic categorization variable of aptitude. It has been suggested that the more a treatment reflects the profile differences gotten from diagnostic aptitude data, the better the chances of ATI arising. Another implementation problem is the logical task analysis of sequence of instruction. Has enough detailed observation taken place of the learning steps that led to the students' failure. It is questioned whether treatments can be built that will result in practically significant differences in achievement sufficiently large to warrant further implementation; perhaps no noticeably different treatments are needed but just massive doses of individualization. The art of prescription writing and implementation is still in quite a primitive stage.

Finally, there is the important topic of prescription evaluation. Assuming one has been able to write a prescription and attempts have been made to implement it, how does one know how effective the prescription has been?

Mann: Special education is characterized by categorization of children as being distinguished from others, as requiring certain types of intervention, and aptitude-treatment interaction issues are critical to the concept of obligation of diagnostic-prescriptive teaching. Glenn, could you bring us up to date on ATI?

Bracht: I might add something that I recently read which, I think, might help us with research in this area—which has largely been hit-and-miss—identifying situations where certain learners are successful and trying to develop alternative programs. We haven't thought too much about what creates aptitude-treatment interactions and I would like to share three models that have been suggested for looking at atpitude-treatment interactions. One is called a remedial model, where the one group of learners is unsuccessful in learning under a typical instructional program and the reason for the cause of that lack of success is that the learners do not contain the necessary prerequisites for the instructional program so some remedial work is to be done. We need to go back and give them the skills that they are lacking...once they have these skills, then presumably they can then learn the program that was successful for the other learners so it is a matter of developing a different instructional program. It is rather a matter of providing the necessary skills to the child, which alone has to learn from it. This is one model and perhaps it is one that might have much usefulness.

Mann: It is a popular model.

Bracht: Another is called the compensatory model. This is one where the instructional program that is effective depends upon certain types of abilities that the learners may, or may not have. But, those learners who are not successful with the program have other abilities and you might utilize these to develop an alternative program; we would be then compensating for the lack of one ability by using a different ability to obtain the same outcome.

Mann: That is another popular model in the field.

Bracht: The third one is one that I don't recall. It is called the preferential model. I think it refers to the students' preferential style for approaching certain things. Some people may approach a certain instructional setting in a certain way. Others approach it in a different way. Maybe some approach it more visually—they tend to learn better through visual processes. Others...it may be through auditory processes and so, it is not a matter of not having certain abilities, but it is a matter of operating in a certain style. So, we develop alternative programs to suit the preferences of the learners. I don't know how useful it is to have these models generate research on aptitude-treatment interaction, but it at least is better than what we have worked from in the past.

Hopkins: Then wouldn't you agree that if there is an area of investigation in special education where there will be treatment/aptitude interaction, this is a very likely environment for it? Certainly if you have blind and deaf kids and you've got an auditory method and you've got a visual method, you are going to get an apitutde-treatment—these populations which have severe disabilities in some particular sensory motor or neurological functions, where conventional materials just don't do what they usually do.

Bracht: I think that one of the problems we have in designing alternative treatments for aptitude-interaction research is that we don't carefully study the processes that learners go through in various situations. I think this type of research has to be done before we can successfully carry on ATI research. We've got to find out what processes the learner does go through when he is in a particular instructional program.

Mann: When you say processes, do you want to elaborate on that a little further because process has a number of implications, in terms of work, with the handicapped. You talk about process variables, or the actual acts of learning, themselves.

Proger: Are you saying, Glenn, as far as a prerequisite to really getting into the aptitude-treatment interaction studies, that we need a lot of descriptive studies of actually what is going on in learning? Some people say that we try to become too sophisticated in educational research and that we are trying to study something for which we don't have the complete tools yet and that we come up with a lot of irrelevancies, as a result of the inferential statistics being used inappropriately.

Bracht: I think that yes, we probably need to go in and do some very close and careful observation of learners—some detailed case studies with individuals, to find out what they do in various learning settings.

Mann: Actual behavior—or inferences therefrom—is this a black-box approach you are talking about, or is it inferential?

Bracht: I am not thinking of an inferential study at this point... I am just going at it as carefully observing learners and using these observations as a basis for designing treatments.

Hopkins: Aren't you saying that people need to develop some kind of theory about how the learning is taking place. How are you using the term, inferentially?

**Bracht:** I think we need to infer, from our observation of learners, what processes take place in learning.

**Hagin:** What sorts of observations are you recommending?

**Bracht:** I think it might involve interviews to a great extent, where the teacher talks to the learner, trying to find out from him what he does in certain learning situations. I think, in other cases, we can observe his processes. What steps does he go through to solve a certain problem? Does he take this approach and find it unsuccessful and then try another approach? What approach does he usually find successful... by following rules, or is it some other approach, that leads him to a successful solution? That type of detailed analysis may be useful. I haven't done this sort of thing, so I don't know.

**Mann:** Let's put it this way. This really isn't conceptualizing processes inside the learner. You are saying, maybe he is confronted with this—and he does this, or this. Is that a process model, or just a descriptive model?

**Proger:** You are still looking at processess—learning processes. You are just describing them correctly.

**Bracht:** It may be observations of his psychomotor processes, as revealed in performance, and from that we would maybe infer what goes on "inside."

**Feldhusen:** I wonder if I can ask you this question. Is there not a good possibility that the kind of observational study you are proposing would lead us both to a better understanding of how to design treatments, as well as a better understanding of how to classify people, for the classification variable? Too much ATI research up to the present (and Cronbach and Snow say this, too) simply has taken test scores off an available test—after the fact—and classified subjects as high and low, on whatever that test score says and then asks "is there interaction?" We need a bigger view of how to classify people. It would seem to me what you are suggesting—observation studies—might lead us both to a better understanding of how to design treatments that really make a difference for certain kinds of people, including the handicapped.

**Bracht:** I think so. I think we perhaps are, or have been in the past, too much bound to traditional ways of classifying people on the basis of what valid tests we had—low ability/high ability—low achievement/high achievement. If we can try to find other ways of classifying people or learners, we might be quite successful.

**Feldhusen:** I think it clearly is an implication, from that Cronbach-Snow monograph, that both the treatment and aptitude thing probably have to be conceptualized in bigger, broader ways. Often our treatments in ATI research have been narrow, little short-term things with the aptitude being fairly precise but too narrow, so that one very gross generalization is that we have to look for bigger, more complex ways of characterizing people and treatments.

Hopkins: I think what you say, John, could be misunderstood. I think one of the things that Glenn noted in his review was, most of the treatments have been very complex, hodgepodge collections of treatments and the learners have been categorized in very—again—global, molar ways, such as IQ score and that it was well delimited and refined and narrowed down treatments and the ways that the learners are categorized that seem to have the most promise for speaking about ATI. I don't think, however, what I am saying is really disagreeing with you; I think you are complaining about small treatments over a couple days—and not really potent ones, or maybe you could clarify it.

Feldhusen: I think that I would like to agree with the point you are making for basic research. But for practical purposes, and it seems that most of ATI people are looking for practical solutions to instructional problems, I expect to emerge not such precise classification as you are suggesting here, but rather global treatments and global classification as being meaningful, but hopefully not ambiguous, in the treatment or the description of the learner. I am not arguing for ambiguity—I am simply saying that I would expect both meaningful aptitudes and treatments to be fairly complex when finally arriving at a stage of application with a group in a school setting. Back at the basic research stage—which we certainly are at in this domain—yet, why I can see much good potential for the alternative you are suggesting—the precise aptitude definitions and precise treatment definitions.

Mann: This is a really critical issue in the work with handicapped because a small population is so variable. To precisely characterize pupils on precise aptitudes means that we have knocked out most of our population or our sample, rather. We have very small samples in much of the research we do, anyway. We begin with very small _Ns_ and we refine them still lower. How would you manage this issue and how far do you generalize, from the small _Ns_ that you end up with? Joe?

French: I was thinking as you were talking, about that very important thing that Rosa started off with,and that is characterizing the kids before the study begins and really knowing what the children—what the sample is like and, of course, one of the real problems in special education is that tests are standardized on the representative samples of the United States as a whole. Then, when we get to a deviant group, we really don't know how well that test applies to those kids and so when we want to generalize from the tests to instruction, we don't know whether the kids in our sample can be generalized, too. I think we need one instrument to make real fine delineations of the characteristics of the learners and the processes by which they learn, and I don't know how you get to that kind of data.

Hopkins: Well, maybe we ought to begin with what Glenn was saying... something like this...parents have been observing their children long before Piaget came along, but he saw things that they didn't see, even though they were taking place, and I think he's suggesting that the really critically-trained, sensitive person looks for things in the child and the educational interaction, and maybe finds some things that other people haven't seen before, and this may help us to group children into meaningful populations.

Bracht: We might consider this approach. Given a defined population of learners, let us ask what is the best program that we presently have for achieving some sort of outcome? We could administer that program and observe which students are not successfully performing and then do a very close observation of those learners to find out why not—try to find some explanations—of performance, and see if there is any common characteristics among those learners, and we, at that point, may identify the remedial need. Not that these learners are really that much different from the others in most stable characteristics. It is just that they are different, in terms of the prerequisite skills needed for this program, and so a remedial treatment may be the best approach. Or, we may find that these learners differ in some stable aptitude and that by having the compensatory alternative treatment, we can obtain successful performance, or some other approach. I think those are the learners that we are going to have to study very carefully—the ones who are not successful with the best program that we have, at present.

Mann: This should characterize special education. By the way, your approach has been used, I think—at least Kirk reports this. A specialist goes to the classroom and watches the youngsters who are failing and keeps adjusting the program and materials until it fits them the best—it emerges, somehow.

Hagin: When we visited the Institutes of Defectology, in Moscow, they told us that every child who was two years behind in normal school progress was referred for help. And we asked about children who do not improve and remain two years behind and they acknowledge there were, and so we said, what then? And then they said, very ominously, "We look into the school." We, then, would look into the parents...the environment...the child's disability. It seems to me, maybe the Russians have a good idea in doing something of what you are saying, Lester, ...in looking into the educational treatment first before attributing educational problems to non-educational causes.

Mann: Well, this is, of course, I think increasingly happening. I think the schools are being told, don't tell me the child is handicapped. You teach him.

French: The Right to Education Decision in Pennsylvania gives evidence to that...that all children are entitled to an education.

Bracht: Perhaps one of the weaknesses of ATI research design is, then, that we tend to think of all learners as being of two or three types: You know...those learners who best learn with treatment A and those learners who best learn with treatment B... and maybe that is not the case. Maybe a great majority of all learners will perform successfully with a particular best-available treatment and those who do not perform successfully with that treatment...they probably have a variety... there may be a variety of explanations for their poor performance. Maybe for some, skill 1 is defective; in others, skill 2 may be lacking. In others, maybe several skills are lacking. So, there is no single remedial treatment that will bring up the performance of all those who are unsuccessful. Maybe ATI enables just an individualized approach with those who are not successful with the best-available program.

Proger: John was saying perhaps the aspects of the interactions are not conceptualized broadly enough, and I believe you were saying earlier that a more general conception is needed. May I ask if you believe that when we measure "aptitude" in an ATI study, to what degree should we attempt to individualize treatments and stop there, or do we seek guidelines for precise aptitude-treatment matches? Will the ATI research tell us what to do along these lines?

Bracht: In trying to express my personal opinion here, I don't think that we are going to find too many generalizable aptitude-treatment and interactions. I tend to agree with Gene Glass about generalizable findings in education. So often, the results we get with experimental studies are context-bound. You know, we find certain results in this study, but maybe it is due to the administrative support that was given for one of the alternative treatments. Or, maybe it is because of the teacher variable that we get a certain outcome. Maybe it is because we have one or two very responsive students in one of the programs. I think, so often, our results are bound to the context in which a particular study was done and we don't then find results which are generalizable to other subjects and across the population. And I think the same may be true with aptitude-treatment interaction—that we won't find much in the way of generalizable results. In one situation there can be an identified aptitude-treatment interaction; some students are successful with one program—

others are not, but might be successful with some alternative treatment. However, I don't think we will be able, personally, to identify very many generalizable interactions once we leave the context of the study.

Feldhusen: It should be noted, of course, if there is an answer to the criticism that Glenn has just stated—not implemented yet, but some of us at least have suggested—that the answer is very obvious to the problem that Glenn has just proposed. If there are these many uncontrolled context elements, then why not go after them and find out what they are and nail them down, too. In this particular, Terry Daney, I think, make recent attempts to do that. For example, with interaction analysis, environmental assessment of the type Barker and others, have used and so on. So that one doesn't have to accept what you just said...although there is not that evidence yet of utter success in nailing down those confounding elements.

Mann: Glenn's part, however, is particularly well taken with the handicapped. Repeatedly, in major projects with the handicapped, the disadvantaged, etc., the efforts, when transplanted, really don't work out. The charisma of the principle investigator who has made the difference, perhaps in some way indefinable, is not available and the initial success doesn't re-emerge, or the project director leaves the original project and it fades to a limited simulacrum of the initial work.

Bracht: John's point is a valid one. I must qualify my statements. I don't think we are yet capable of finding generalizable interactions. I think if we are, in being successful in doing the thing you describe, then we might be able to find them.

French: If we were able to characterize those who did succeed in program A in much greater detail than we have ever attempted to do before, I think that might be the first start, because indeed, program A does work with some and then how do those with whom it works differ from the others? One of the things we have not mentioned would be personality dynamics. I see an awful lot of kids who might be called passive resistors; they've got all the cognitive characteristics that are necessary for success, but for some reason the program doesn't work for them. I guess it is easy to say it's because of personality problems, but...I think that is a dimension we have to look at, as well.

Hopkins: I think there is a peculiarity in our ATI discussion here. On the one hand, we have difficulty in finding treatments to interact with people characteristics. Yet, then some of us are saying that treatments interact strongly with contextual variables, or more explicitly, that there are treatment-by-treatment interactions. So, it is kind of interesting on the one hand, that we are able to find so little evidence of personalogical variables that interact with treatments, and yet we say that the level of generalizability of ATI is due to the power of these variables.

Bracht: In reality, the "real" interactions are probably four-factor, five-factor or six-factor interactions, but we've been trying to find two-factor interactions and I think we haven't been able to find those two-factor interactions because of those other interactions that exist.

Mann: The problem, then, is perhaps too many, rather than too few interactions. Jay, you have been sitting there recovering from your Mohawk flight. IT IS TIME to have the child of wisdom speak up!

French: While he is preparing, let me stick one thing in here. Bernie House had a real interesting title for a monograph that was published in the Illinois gifted program evaluation project, entitled "Advocacy in a Non-National System." And the conclusion was, that the intervention programs that were successful were the ones that had a strong advocate in the system and I guess this is what Les was referring to as the charismatic leader...that a guy really wants to make a thing work is very instrumental.

Millman: I believe...we have hit on a number of topics and for my own piece of mind, I need a little closure of it. Let me go to my notes and pick off some of the topics...perhaps starting from the back and working forward. My most current notes are on the ATI. On the ATI business, perhaps we need to not only look at the context and look at the learners, but we have to look also at the nature of the treatment. Joe, your colleague at Penn State, Joe DiVesta, has just completed a major ATI study and I believe his findings was that the things that interact with a particular treatment are things that are very close to what the treatments are themselves. It might be that we need to study not only the "process" but also, more closely study the structure of the subject matter in ATI. My earlier notes concern what is special about special education and I was trying to get going on what kind of problems there are in special education and from what I have heard so far, it seems that there is particular concern with the evaluation of instructional materials. And by the way, I would like to make a distinction, I think, between instruction and curriculum. It seems to me, what is being talked about is not curriculum evaluation, but evaluation of instructional materials. Evaluation of curriculum is quite a different thing and it involves values; it involves a whole host of considerations that usually don't get considered when all you are talking about is instruction...I think what is being talked about here is instructional material evaluation, not

real curriculum evaluation. It seems you special educators are also criticizing testing approaches, especially whether some are appropriate for some groups.
Then what is it possible to teach a child at a given age, and how do you define relevant characteristics of the child for specialized instruction? Now these, of course, are questions that could define fields other than special education. I am wondering...are there other kinds of questions that people in special education are particularly concerned about.

Mann: You have hit on a big issue. Certainly, Jay, there are many other problems that are unique to special education—not unique, but perhaps salient to special education given small number of youngsters—populations and students characterized by medical diagnosis, noted educational characteristics, upon standardized IQ and ability tests, to place children and determine instruction, etc. ATI, of course, is one of the dreams that all of us working with handicapped hold. Can we define a materials or programs or curriculum or whatever, for a special pupil with special characteristics?

Feldhusen: It is nice to know that Jay mentioned the DiVesta study which shows some rays of hopefulness for ATI and that, in spite of seeming gloom otherwise, there are other rays of hope—notably in a very fine monograph from David Hunt at the Ontario Institute Study for Education, in which he argues very close to what has been suggested here—that if treatments are designed to fit with a good understanding of characteristics of the learner, which hopefully grow out of the kind of observational study, perhaps that Glenn has suggested, but then one may hope for significant interactions to occur. Now, Mr. Hunt operates within a total theoretical framework, too. It is not just a hit-and-miss looking for treatments and a hit-and-miss looking for characteristics. In that monograph, he suggests that there has been eminent success in fitting or designing treatments that really make a difference for particular subgroups. These do not include special education groups, however.

Mann: Can you give us an example, John?

Feldhusen: Well, you know of course, that that group of people work with the information processing model that is attributable to Harvey, and others. Their method of characterizing the learner has much to do with his cognitive style and so their treatments are broad, relatively ambiguous definitions which they simply try to get a teacher to implement. Cognitive style would be a complexity variable with prior assessments being made of a group of children's higher, lower, medium status, or whatever. They have successfully devised ways to get the teacher to adapt her methods within the room to subgroups of children on the basis of cognitive or intellectual level and cognitive complexity and seemingly have, in at least five or six studies, demonstrated differential success. That is, they were able to differentiate a treatment that works best for specific groups. I think it is very promising.

Millman: I am just wondering whether we are not trying to pin down factors that might make a difference, but really only a very small difference in terms of the education of the child. Obviously, in this world there are single causes of learning problems and we will direct training to this. Everything is related to everything else, but I wonder when we look for aptitude-treatment interactions that might make a miniscule difference, at best, as to what really is the appropriate research strategy, perhaps we might be just better off to concentrate on things that might make a quantum difference, in terms of what happens to the kids. I am just wondering about the experience of the special educators here. Are there any things that are very important things to do because they really make big differences when you deal with certain kinds of kids? For example, is it true that if you have a kid who has one of a thousand different syndromes—does it really make a difference to put him in a room that has few distractors—is it that kind of thing that you recommend? Are there examples, in special education, where there is something that is so important that you know what to do with this kind of kid; where it is clear this is how the education ought to go?

Mann: You picked a bad example because the distractors or the cubicle approach to limit this has come loudly, but gone very quietly over the past years and I don't think the issue really was resolved, by the way...it disappeared...which is true of a lot of educational issues—research issues.

Millman: I'm just beginning to wonder whether really, when one focuses on a kid who is very low IQ really, you know. Should you do dramatically different things with him, or just take the pace slower? Are there examples?

Mann: Well, there are thousands of research interventions and many reach statistical significance. The question, however, is what does it mean when you reach statistical significance—does this make a real difference? By and large, not. But if you are asking are there legitimately different educational needs and goals for certain types of handicapped populations, I would say, yes.

Feldhusen: One of the problems discussed in a little paper of mine is that long range research—I talk here specifically about longitudinal research—is extremely difficult even today, after all these years of diverse and large-scale funding. And this is the kind of research we really need to have in order to pin down practical treatment effects. It's extremely difficult to find an agency that will stick with you through the many stages you must go through, from reaching the first significant differences in an experiment through the subsequent stages of finally making a difference in school. Hopefully, the regional labs and the RD centers have come close to this, but still, for most large amounts of effort, one can expect termination after two - three years of a project, and then the agency is no longer interested. Their priorities have changed and one just can't go on. This is a significant problem.

Mann: Also, the investigators change. It is hard to keep a staff past a two- or three-year period.

Hagin: I think if you were really interested in longitudinal work and you, the researcher, remain in a setting, it is possible to do this but one must be satisfied with low Ns. I am thinking of Barbara Fish, for instance, who is in child psychiatry. As a young psychiatrist, she identified children of schizophrenics born in institutions and she has followed them for ten years. Now, there are not many; there are probably twenty, but she knows so very much about the life-style of these children, the kinds of manipulations that seemed to have worked, etc. Now this is not the kind of thing that will lend itself to large, statistical treatment. I think you have to be satisfied, if there is to be depth and long-term study, with possibly less in the way of quantitative treatment. I personally have been involved in blind evaluation of the psychological status of these youngsters and it has been most interesting. I have learned a good deal about heredity, seeing these ten whom I saw. Individual investigators who remain with the same population have the opportunity to do informal, longitudinal studies. However, one must give up certain assumptions about statistics if one is to have this kind of material available.

Millman: I haven't thought this through, but I think there may be a distinction that can be made here between this kind of research endeavor, which is part of a longitudinal, clinical kind of approach, and the engineering, technological kind of approach that I would suggest, might be the more fruitful one in special education. I don't know how far we are going to get by looking at the, you know, learning about a few kids very well and all of a sudden you hope to find something from this kind of investigation that will effect thousands of children. In other words, I guess it is kind of a conception I have of the world that leads me to believe that for every child there probably isn't a specific kind of instruction or treatment that is going to be so much better for him over, say, an average kind of treatment. This is, I think most kids will probably do almost as well with, you know, the same particular approach and that might be the better way to go in research in special education, than to try to maximize the yield through that little bit more that you might get by individualizing special treatments for each person.

Proger: Group instructions back in vogue?

Millman: Perhaps pacing will account for enough individualizing for almost all kids. Obviously, if you have a kid who is blind ...I am not suggesting that you ignore the fact that he is blind, but I am saying look all kids who are blind...can't they sort of be treated in the same way? Now, I realize that among those blind kids there are all kinds, you know, they are not all alike—no two people are alike—but just because kids are different, does that mean that we should try to find a separate treatment for each one? Maybe the difficulty that special education and general education people have, in coming up with things that make a difference like, really, just reinforces the fact that maybe the best kind of special education might be just give all children in a particular group essentially the same thing.

Mann: Well, that has been our premise, you know, up in Montgomery County. You know we say that the curriculum applies without concern for modality information, aptitudes, etc., whatever, that we will do the job, by and large, if we pace the children. So far, at least in a one-year study, it has been working. Every youngster has progressed...some slower, some faster, but everyone has gone through a standard curriculum with a mastery learning approach.

French: An example. I had an occasion to visit a school down here. It has about forty severely handicapped/physically handicapped kids who are cerebral palsied. Some of the bright ones learn a lot more than others and some of the duller ones learn a lot more than others, and it is pretty difficult...these kids are maybe five to ten years of age...forty of them...and I don't know...the standard curriculum helped some of them, but then there are maybe half of them that you say, "gee, there really ought to be more progress than that" and then the question is why? and I don't have an answer for them.

Mann: We won't argue that there are bad curricula, Joe, and that good ones are not often applied properly. But, we believe that a curricula approach can do it—properly paced and monitored.

French: I've seen a kid there for three years, who makes apparently no academic progress, yet he is a very bright kid.

Mann: We've seen this in our own schools—we have a clinical education project. Many of the youngsters who had flunked school were referred, and we had a teacher go in and work with them and, lo and behold, they all learned rapidly with individual attention. This may be another key—individualize the same instruction. Don't change the instructional approach, but individualize it, and somehow you get this match-up. Certainly Bloom and Carroll say this.

Feldhusen: I think I've had the complete and final answer to this problem for a long time. And it represents a great faith in individualization of instruction, while maintaining much agreement with the pessimism that Jay just expressed here, too. I think that the work that I was able to do under Mr. Klausmeier's direction a long time ago when I was working on my doctorate, and a number of things I've done since, have convinced me that the most relevant dimension to look for in individualizing instruction is the youngster's just-prior knowledge, or knowledge up to that point, about whatever you propose to teach him next. Now, what we first demonstrated in those studies in Wisconsin was that if you would find a way to diagnose, not a host of personality and other disorders (and incidentally, we were working with mentally retarded educable youngsters at that time), that if you would find a way to assess as precisely as possible, how much the child knew about, say an arithmetic task you now propose to teach him...(we simply logically conceptualized what we thought he ought to know up to that point)...we could take him into the next stage. We tried this out over a period of a year with these youngsters. We did, incidentally, very long-range retention and transfer studies. Our retention studies ran up to something like six to eight months and what have you, so that we could produce very durable, well retained learning. That work supported the Gagne hierarchy concept, which as a general proposition, is a very sound one. That is, if we look specifically to the things related to what we want to teach, we may make quite substantial differences occur when we go on to the next stage, or in producing new learning.

Millman: I would, of course, agree and that goes back to an earlier comment that we should look at what it is we are trying to teach—what the particular learning task is—to get clues for appropriate instruction.

Mann: You aren't referring to aptitudes, necessarily, but rather instructional stages, as it were.

Feldhusen: We'll keep specific aptitudes if you'd like, but they are really not likely to lead to interactions. They are apt to lead to across-the-board effects of an utter, or gross, type of a, or total type of individualization. But, aside from research efforts, I do spend some time trying to observe teachers and I judge that it is a characteristic of the good teachers that they make substantial efforts to try constantly to have at their fingertips the best assessment they can, of where a child is in reading—what he knows up to this point—what his deficiencies are at that point, and this as an antecedent before the attempt to teach the next level or the next stage.

French: I've often wondered what special education is, and I keep asking that in doctorate orals, in hopes that somebody will tell me that...so I can understand. But, at least some people say it is programming for children that are so deviant they can't be handled in regular classrooms, and I think what we have been talking about a bit today, is those individuals who do make regular progress or at least are expected to make regular progress.

Feldhusen: Well...I said nothing about the setting, or the methodology in which, or how one implements this finding of the appropriate level. I do happen to have considerable faith in semi-programmed instructional material as a way of accomplishing this teacher-directed, good programmed instruction. I am much impressed with programmed instruction in reading.

Hagin: Do you think the reason this is more productive is that it eliminates some of the teacher variance? You know, there was an interesting study which Jean Shaw did years ago on first grade reading. When she visited classrooms, she found that what was taking place in the classroom was very much different from what the curriculum director in the classroom thought was going on in the classroom. Now possibly, programmed instruction, by cutting down the amount of variation among the individual teachers, may get a particular program across.

Mann: The research generally agrees that the formal curricula becomes dramatically altered when applied by different teachers.

Feldhusen: I'd accept that as a possible explanation for part of the success of prepared-programmed type instructional materials. When I use the term programmed, however, I don't want to limit the definition to the very simplified linear approaches but am, rather, referring to something that is sequentially organized. But, my major point was that I think the whole implication of that research I referred to is that if the gap is substantial between where the child is in his knowledge of—I'll come back to arithmetic skills and the level from where the prevailing level of instruction attempts to go on—that if that gap is substantial, that he simply cannot learn. That is the problem that besets a large portion of the aggressive, disruptive youngsters whom I have been studying these very many, many years. The gap grows and instruction becomes increasingly meaningless for these youngsters. At the second or third grade level, it is already a year and a half or two retarded, and yet, they move up—the idea is really simple-minded. If one attempts to teach you way up here, and your basic skills are back there, you are in trouble and that, I think, is much more important than the mere fact that the programmed instructional material removes the teacher variable. But, that is just a guess.

Hopkins: I was going to make the observation that this, of course, is an aptitude-treatment interaction when the pacing requests are different for different children.

Feldhusen: You will have to explain. I am not sure I follow you.

Hopkins: You're talking about individualization. You are talking about applying the treatment at a different rate, or at a different level, so that, in effect, you can say there are different traits operating.

Feldhusen: Level was all that I was talking about for the moment, but rate enters. As a matter of fact, the research that I described that I was lucky enough to do with Mr. Klausmeier, rate did enter there...fifteen or sixteen years ago we also taught for mastery and that forced a rate upon it. That is, we would not—could not—move on, or leave the child, until he had mastered the task. So, there were two elements. It was both focusing in at what we believed to be the appropriate level, and then teaching for mastery at that level, and that did force rate upon us because it might force us to spend more time with a given youngster in order to achieve mastery than with another youngster.

Bracht: The research that we traditionally have in ATI has been of the design where we develop alternative treatments, and they have been specific treatments...not too complex. The directions have been clearly specified for administering the treatments, and then we observe the effect of these treatments by administering some dependent variable. I think we need to supplement this design when we are doing research on instructional programs (which we, today, are more likely to call evaluation) because the treatments can be, and usually are, administered in different ways by those who are responsible for administering them. Also, the treatments may contain a set of very complex things. Some may be emphasized more than others or some are administered more effectively than others. So, I think we need to supplement our traditional ATI research design by carrying on some very systematic observations when the treatment is administered, so that we can identify what components of the treatment are being administered as intended: what components seem to be effective; try to find explanations for the performance of others when the treatment is administered. I think we need to use some of the techniques that have been advocated recently for evaluation when we are carrying on research on instructional programs, rather than just administering treatments and observing and giving a post test.

Proger: John, you were saying about longitudinal studies...that you would like to see this done wherever possible. How do you feel about the short term in-and-out, one-day studies where you investigate some psychological process in a child. For example, one area that I have been particularly interested in is Rothkopf and Frase's studies of prose-learning...how you can organize a child's thoughts, asking certain questions in certain positions throughout a passage to keep his attention moving along. And those are all fast, in-and-out studies. You can run them off in an hour or something, on one passage. Yet, you do not know what the long-term effects of these are. How much faith can we put in the results of such studies? Sometimes you get some beautiful differences; you can get interactions—ATI's running all over the place, if you try hard enough. How generalizable can we be with the results of the one day's study? Yet, we see them all over the place.

Feldhusen: I tried a couple of rounds of prose-learning and came up with an unexpected result. On the second round, in a sample of about twenty-five college students, there was powerful protest against my machinations; they knew what was coming—another round of this. I decided to talk it over with them. The arbitrariness of having to sit and read something, and the fact that somebody expects you to learn from it while you sit there reading it, came home to me. Like, if I said to all of you, now read the first page of this, and the second time around you knew there is going to be a test on it, too, and knew it is apparently a very obnoxious kind of thing. My experience made me very suspicious of problems of generalizing, from this kind of activity, or, just as I think Frase and Rothkopf's subjects were always paid. Yes?

Proger: Two bucks...five bucks...something like that.

Feldhusen: So, of course, if Les offered us—each of us—$20.00 if we would do it, why maybe we would be more receptive here, too, to risking our egos and what have you, doing this sort of thing. But, well, I think that I am generally saying that I am pessimistic about the sort of work, because as I look back to textbooks that I use with questions seeded into the text, and/or the other types of organizers that may appear there, that the little I know about characteristic learning styles and teaching suggests to me that there may be a tremendous gap between laboratory-oriented analysis of prose-learning behavior and the student given a textbook and told to go off by himself and learn. Our students at Purdue and, I guess, everywhere else...their style includes felt-tip pen with heavy underlining, in preparation for objective tests, and a lot of other characteristics that are quite different. So, I guess I've answered your question by analogy and indirectly. Let me just add that one has to start somewhere and I think that it—yours—is a very promising effort...I think there is still much of value that may come out of our understanding of learning from prose.

Hopkins: I would just say isn't that just a special case of the problem we have in going from experimental psychological research to actual educational practice? I agree with John that we need people doing it as a first step. We also need to be well aware of the risk involved and not be naive in assuming that we are going to get the same kind of results when we seek them in a classroom.

Proger: The thing that sort of hangs me up when I read that stuff and we've tried a few studies of our own along that line with the prose-learning and some of them bomb out; some give us some results back. Something like that has such big implications for the construction of learning materials...individualized program instruction materials—it is not much different really. You talk about inserting questions and continuously monitoring and branching kids around. You could easily twist it into something like that, but without a long-term study...day-by-day...you know, just grinding along to see how long the kids hold up under that—I don't think we really have any evidence at all.

Bracht: I think too many of us, also, are building on the faith that if we can identify various specific manipulations which are generalizable and can then build better instructional programs —the complex things. I don't think we can show a lot of success in that area, and there has been some suggestion that we ought to go the other way. Build complex programs and begin with them—do some very careful study of them—try to determine after our observation, what things seem to be working and from that we will get some really generalizable observations of things that work.

Millman: That was Bart's initial question.

Feldhusen: In relation to the point that Glenn just made, there is this very interesting study in progress now...I think AIR has the contract through the Office of Education to trace the complete...hopefully, the complete history...from, supposedly from, very basic research to the current widespread field application of thirty substantial educational innovations like Project PLAN...like the Schutz Reading. Only one volume of this is done, and that is the one on Dick Schutz Reading Program, but the volume attempts to trace back all the way through to even his early involvement in some of the basic research on programmed instruction and stage-by-stage how did we finally end up with the thing that is now, I guess, being implemented on a complete statewide basis in a number of states. Now, that list of the thirty educational innovations that have made it is very interesting. They intend to produce a monograph for each of them, trying to trace the complete course from very basic research to widespread field application. We might learn whether there really is hope in this sort of continuum, as a result of these monographs.

Mann: Perhaps we can begin the second part of the discussion with this provocative point Jay raised offstage to me. Can you repeat it, Jay? It is essentially an elaboration of a point he made earlier.

Millman: I guess what I was trying to say is that I have a faith...a belief that some things are going to be more effective in educating children than others. There is no question about that ...that I think proper pacing is important in educating children. I think it is important to look at the requirements of the learning task before you go off to try to teach. These things are clear. What is not so clear to me is that there is an interaction between the various ways that you teach kids and the kind of kids that they are. That is, I have a sort of a hunch that if what will work well in teaching learning-task A to a mentally retarded child, will probably also work well for a child of normal ability and vice-versa. Now, there are obviously some exceptions where you have some gross physiological problems, such as blindness or deafness, where obviously certain kinds of instructional approaches are not going to work, but for the general kind of situation, I am just wondering if there is, indeed, the case that where you've got a mentally retarded kid you've just got to do something completely different in order to educate him effectively than if he was of normal intelligence. And you know, it seems to me that perhaps, after all these years, that the burden of truth should lie on those who argue just the opposite—who argue that looking at the defects of a child, or by doing intensive case study of kids, that this will help us to educate them better. I would say that the burden of proof is on those who argue that way. And then, I might close by raising a question. Can you give me examples of approaches that are very effective with certain kinds of handicapped kids that would not be effective with normal kids who would not be classified as handicapped?

Proger: Along those lines, you might be interested in the Texas state education department's efforts which goes very, very specifically into social work, histories, and genetic abnormalities, etc., to whether—when you know something about that—whether it applies educationally; does it really mean a heck of a lot of difference for the kid in the long run. This is a huge study that has several thousand kids, trying out different diagnostic approaches with different types of professionals, and seeing whether the addition of different types of information, and then programming for them, makes a difference. That study is currently under way in Houston headquarters of their regional state office. It is a fairly tight research design, too.

French: I think that one of the problems that create difficulties in research with the handicapped is that most of the handicapped kids are in regular classes, and it is usually only those that have become so deviant that they have had to be separated, that are studied. So, we seldom include those not-too deviant kids and those with milder handicaps in our studies. The administrative arrangement is to put those who are most deviant into institutions or into special classes. So, I certainly wouldn't generally disagree with the statement that there are educational principles that apply to all kids including the handicapped, and I guess those that are in special education classes or in programs, are those that are not learning by the most popular methods.

Mann: They flunk school, essentially you are saying, and that therefore, they are placed in categories as handicapped.

French: I was studying in this high school drop-out study...these were lots of kids with IQs of record group test scores of 110 and above and a number of them, with IQs of 130 and above. They left that system by choice, usually.

Hopkins: I think, in line with Jay's point, a couple of anecdotal situations come to mind. I am a drop-out school psychologist and worked in a district where there are a half-dozen school psychologists, and among ourselves we would mention—and kind of joke—but, there is a lot of truth to it—that it didn't matter what we found out in the assessment we were there to rubber-stamp the kids. The rubber-stamp was educational prescription for the teacher to use. It really didn't matter whether there was emotional disturbance, or suspected neurological impairment, or whatever—the educational strategy recommended was essentially the same, irrespective of diagnosis. The other point, and of course, that reflects on me as a school psychologist—I didn't know enough to be able to provide a legitimate prescription. In any case, I would support Jay's position. Diagnosis was of no help at all, in terms of educational prescription. If then, diagnosis is irrelevant except for status, John's point of working with where a kid is, irrespective of cause, and then taking it from there, seems a sound position.

Millman: This isn't Monroe County!

Hagin: Well, I think Jay's point emphasizes again the embarrassing findings of about ten years ago, when Orval Johnson assembled all the results of special education...the effectiveness of special education. I think the title was, "What's Special About Special Education?" and there really wasn't very much, and I think the next step is John's. The kids who eventually get in special education are those for whom the regular, standard program, whatever that is, doesn't work. But then, we find special programs don't accomplish much either. Incidentally and parenthetically, in New York City I get a sense of history when I go into the schools because I can see everything going on from the very best and newest, to what was current in my father's lifetime—going on at the same time. So, I don't know what is a standard approach in school. The kids in special education, in any case, are kids who are not going to learn any place else and maybe with that comes something besides just individual handicap with the sum total of several handicaps. It is like kids don't get sent to Bellevue if they are going to get better any place else and so you work with this. We have what we call a typical Bellevue type and it is something different from just a plain ordinary schizophrenic from West Chester County.

Mann: Rosa raises the point about what's special in special education. Well, if you put a person in the hospital and the hospital is lousy, physically, and he gets no care—no medical attention—the hospital won't make him any better. Presumably, the hospital is a place where you focus treatment, or training, or whatever, on somebody. Presumably, the special class is a place you focus training techniques on children with special needs, and this may be what defines special education. If no real help or if cursory help is given, the children certainly will not improve.

Hagin: I also think there is another important issue now, in terms of the reaction of the minority groups to testing and diagnosis. They are just not going to let school psychologists get their hands on many kids unless they are sure that something better is going to be done for them, as a consequence of that intervention. Maybe special education for this reason will improve, and really do something special.

Mann: We talked about aptitude, and this leads us into one form of aspect—the idea of abilities and disabilities, process variables—as opposed to S and R variables here. It has been presumed that there are basic learning skills or abilities that have to be trained, remediated, corrected or compensated for, through educational school learning or with the help of school learning. Rosa and Archie Silver have been very active along these lines.

Hagin: Well, as a result of the follow-up studies that were done in 1961, we began talking about deficits and teaching toward deficits, rather than teaching along the lines of compensatory model. This is applied in several of our school projects, and in regular schools, working with teachers, hopefully, so that some of the children will have less special kinds of problems can be worked within the classroom. We are finding that in intact, "normal" groups, thirty percent of the children require some kind of an educational intervention. It is unrealistic to think in terms of special classes for these or...we must work with the mainstream of education. So, what we try to do is to keep the child in a regular classroom, insofar as possible. Sometimes we run into a child who must be hospitalized. This comes maybe one out of 100, but many, many children can be worked with in regular grades and finding deficits does not necessarily mean removing the child from regular education.

Mann: You are talking about deficits. Are you referring to educational ones, or behavioral ones, or process ones, or whatever?

Hagin: Usually, combinations. For example, we are interested in teaching reading and we think in terms of perceptual and cognitive abilities, which underly reading.

French: And so, if they have a visual perceptual deficit, then you would give them visual percetpual training—is that it?

Hagin: Not training. We call it perceptual stimulation. We might run into a child who speaks only Spanish and has very limited use of English. So, this child will receive intervention for his deficits through language stimulation.

Bracht: One approach that might be followed here, is to carry out a detailed task analysis along Gagne lines. We can perhaps construct a hierarchy among certain tasks and identify what prerequisites are needed for each, so that then we can use criterion-referenced measurement to determine if the student has a specific skill, and if he has, then make decisions if he is ready to go on into the next level in the program. That is certainly a monstrous task. It takes a lot of work to do this sort of thing, but I think it would facilitate the outcome of reading, let us say.

Mann: In the learning disability, are there one or two camps? Rosa is in one; I am in the other. Rosa suggests the basic learning skills she is referring to should be remediated in her case. In the other camp, like Bontner, Sabatini and me, we would suggest that they be compensated for. I say that you focus directly on instruction of the area being taught through an instructional objective—another referenced measurement approach.

Bracht: Are the abilities and disabilities you are talking about very specific things?

Mann: They are process skills—as opposed to the instructional skills—I'd say...perceptual cognitive skills...auditory decoding skills.

Millman: That relates to the question I was going to ask...let's suppose in giving a test or something, we notice that a child doesn't do very well in block design..can't seem to reproduce designs...something like that. Would this be the kind of thing that you might then give this child...to practice on...tiles and things...that kind of thing?

French: I think that is what is done. One thing that has bothered me for years is the reliability of that one five-minute assessment of how the kid manipulated those particular blocks.

Millman: Well, assuming that the child then was given more tests and couldn't do them very well—would this be the kind of thing, Rosa, that you might do?

Hagin: Let me say, we are not teaching children to do better Bender Gestalts, or better block designs.

Millman: Neither would I, but I am trying to set up a hornet's nest. I guess...did you say when we find a deficit, what we do is look at the deficit. Now, is this the kind of deficit that you would be apt to find, and if so, is this one that you would look at?

Hagin: Well, we don't use block design. Ask Dr. Wechsler some day what is block design, or any of the other subtests of the WISC. They are very complicated. Many, many functions go into their performance. We try to look at much more simple aspects of the child's functioning. We are interested in Bender because it does relate to school learning, such as penmanship, and so the Bender might be the instrument that we use to look for the deficit and we might train the child with chalkboard techniques, for instance, which we hope would be input towards his ability to write. We think this is the way to do the school work. We think auditory discrimination makes a contribution to reading, and...we have not been satisfied with any of the tests in print, so we have some techniques that we have developed...amateurish, little things that get at auditory discrimination problems, and we see that training in auditory discrimination contributes to reading and all of these deficits are looked at—not just as individual things that must be taught—but they are looked at in terms of their relationship to reading.

Millman: Right. And that is essentially the point that John was making...that when you are trying to teach somebody a skill like reading, you look at what I've been...components, skills, and prerequisites, etc., and required entry behaviors, to use some of the jargon. And you, as I understand it, you look at reading and you see that it requires certain auditory discriminations; you look at penmanship and you see that it requires a certain kind of perceptual discrimination, ability, or something, and you see they don't have that, and I think John would say that is the way to do it.

Proger: If you say that you are for individualized instruction, that is fine, or looking for learning preferences, but I don't think there is enough research to say we teach to the deficit or the strength. Kirk would say, all right, we teach to the strength. We compensate, I guess, in his 1971 book. He states that quite explicitly, and Rosa was thinking along John's lines, Jay. Entry behaviors and ability training aren't synonymous.

French: It seems to me, one problem is with the reliability of the assessments that we are doing. As Rosa said, these are rather amateurish things. I imagine there are a lot more... other amateurish things...that other people are using to make the same kind of judgments, and if your judgments don't have reliability, perhaps they aren't very valid either.

Hopkins: Returning to Jay's comments about Rosa's position, do these compensatory treatments bring about better perceptual performance? That is a very clean, sensible question. And then, the next step is, if they do, does this spill over... generalize into handwriting? But, I think, that is quite different than what John was saying earlier, Jay. I think John is talking about working on handwriting directly, and improving it, by bringing it up to a certain point. John's position is much less theoretical.

Feldhusen: I don't want to be too severely delimiting though...Jay wisely broadened my original position to all the relevant...you know, related skills, abilities, factors which could be established as related to teaching handwriting, so I really wouldn't want my position to be narrowly conceived, the way I first stated it.

Proger: I liked it the first time around.

Millman: Well, I think that there is a continuum. A particular skill almost duplicates on the one end of the criterian task and at some other end, in the continuum of training, is far removed from the final performance.

Mann: That is a useful clarification. John, you originally were saying that when you approximate handwriting, you begin with those behaviors that are specific to handwriting no matter how primitive or simple they may be and gradually move into formal handwriting. That is different than training an ability or basic learning skill, if you will. No...Rosa dosen't quite respond to this model.

Millman: Well, I just made the assumption and I may be wrong, that when she uses the Bender test it is with the assumption that if you can't copy certain kinds of forms, how in heck can you write words correctly. To me, that seems to be right up close to the thing at hand that you want to teach.

Mann: Well, if that is what she is concerned with, we wouldn't have to call it perceptual stimulation, but I don't quite believe that is her theoretical position, which really is deficit training.

Hagin: Well, in 1961, we had the very good fortune of inviting back forty-one children whom I had taught and Archie Silver had examined ten years before, and this was a very sobering experience. We invited these youngsters back—aged sixteen, seventeen, eighteen, nineteen, twenty—and we were impressed with the relationship between improvement on the perceptual measures and reading. Inadequate readers as young adults, seven were those whose perceptual picture was essentially unchanged. I must say the psychiatric problems were not impressive as they grew up. In any case, it appeared reasonable to teach toward those deficits, and we set about to do this.

Proger: The correlation you found, of course, does not establish the cause and effect. So training what you found deficient or diagnosed doesn't necessarily follow.

Hagin: We have done it. We have found that it is not difficult to modify perception.

Proger: But, does that help academic functioning?

Hagin: Yes...sure. And, the changes came. That was interesting. We used different treatments. One was perceptual stimulation and the other was Scott Foresman-taught, according to the manual, as we thought this was the least perceptually oriented of the reading systems we examined. We found that the children in the experimental group managed to maintain the level without any formal instruction from our people. And furthermore, at the end of the experiment, the children who improved in perception improved in reading comprehension, and there really were no differences between this group and the "ready" group in their word-attack skills. It apparently did not matter if one worked with Scott Foresman or whether one worked in perceptual materials, to change the ability of the attack words.

Feldhusen: They improved in comprehension?

Proger: I wouldn't see it would do that, in terms of specific treatment effects.

Feldhusen: Of course, it is discouraging, simply to hear that one group improves more than another, since statistically significant differences may be practically insignificant. Should we not set some kind of a goal that is realistically defined, or can be defended realistically, for expected movement? Do we have to just settle for .05 significance, in terms of performance, on a test? Shouldn't we rather have a definition of expected improvement, in terms of something realistic or meaningful in the child's life in or out of school?

Hagin: We once tried expectancies, in terms of remedial reading, for children—what we used was expectancy-based individual intelligence tests, so that we were shooting toward a goal for each child, in terms of what he could do on the IQ test.

Hopkins: That sort of study appears to be a correlational study, where you looked at those who grew most and compared them with those who didn't, and I am really at a kind of loss to know how solid such evidence is on whether treatments have an effect in reading, and I don't think it is that fruitful to go into the specifics of such a study.

Feldhusen: I think the issue is very important though, that we do, some day, try to assess our tendencies to make judgments from simple, comparative studies which do not address themselves at all to practical concern with the amount of change or improvement that relates meaningfully to some kind of in-school or in-life behaviors. This is a great problem with mastery teaching. Where one is still so often bound here to very arbitrarily defined levels of performance that no one yet can relate meaningfully to what is valuable and necessary for a child's continued functioning in school. Someone might arbitrarily set 90% as his test-performance level in a mastery testing situation, and yet he might be hard pressed to demonstrate that setting the level 60 doesn't produce children who go on and do just as well as those passing at 90. I hope that we will increasingly concern ourselves with the quest for meaningful performance levels setting. This is not to necessarily put ourselves out of business completely in comparative studies. We can still work within that framework, and this is certainly one of the models that Cronbach and Snow suggest, too. Both, if I understand correctly, recommend comparative studies, and yet, ones in which you begin to try to specify meaningful performance levels.

Bracht: One reason why it is so difficult to set standards for tests we traditionally use is because the scores obtained from them are not directly interpretable, in terms of what a person can do. So, if a person gets a score of 20, it really doesn't convey any information about what a person can do. It might give us an idea of—if he is a good performer or if he is a poor performer—or some general classification of that type, but does it say what he can do and what he cannot do, or what kind of a person with a score of 30 can do and what can he not do better than the 20 scores. It is also likely that several people with scores of 30 obtain that score in different ways. They don't all—or can't all—perform the same things or perform them at least to the same degree. I think that is one of the problems we have. Our tests generally do not allow direct interpretation.

Proger: I would like to keep on trail of mastery levels or criterion levels. We agree that many combinations of different arithmetic problems are needed to actually assess mastery of a particular arithmetic operation. The research might say, well, it is sort of ridiculous to consider all those combinations, so why don't we think of sampling? But, sampling problems can really foul us up in research design in special ed, in a way which is all tied in with the mastery assessment idea. These sampling problems are particularly troublesome with the educable retarded of different ability levels. Groups within fluctuation in, let's say, in IQs or any other types of abilities. The range and heterogeneity make a big difference in what happens in their performances or any generalizations from it, and that is of sampling problems I think we would have to worry about, in terms of designing experiments with treatments; or let's say, we take the so-called learning disabled who have a lot of perceptual problems like reversing letter "a" or, you know, rotating it to make it into letter "e" and if you take "6's" and make them into "9's", etc., what then? You get into item sampling. If you don't want to test every last item, what does this imply for our handicapped populations? We can say, for normal populations, that we are going to sample the bodies to represent certain populations. But, using sampling strategies which are in vogue right now...if these are perceptual hangups, and you sample the wrong items, you may have selected the very ones that will throw an LD kid right out the window. He may really know the task, but it is just his perceptual hangups that you are sampling, and would catch.

Bracht: If we would take the skill of adding two one-digit numbers, there would be the possibility of ninety different tests... we could randomly sample from that set of tests. If we think that there is a reversal problem, we might identify those digits that are subject to reversal problems—like 6 and 9. We might make our tests even more specific and include in our item-sample all those pairs of digits where 6 or 9 is one of the digits. Could we even be more specific and focus in on 8 and 3? All two-digit problems involving 8's and 3's...5 and 2 is maybe another grouper where there is some problem, so I think we could define a more specific test domain for certain content areas.

French: Weren't you saying that it wasn't feasible to do that—earlier?

Bracht: If we can define the domain...

Mann: Of critical items...

Proger: This is something to consider in our domain of testing—sampling.

Bracht: And, maybe we don't want to define the domain as specifically as adding two one-digit numbers. Maybe we want it to be adding three one-digit numbers or adding three-digit numbers—less than twenty maybe, which would still get at the essentials of adding two one-digit numbers.

Mann: Jay? You were shaking your head.

Millman: Well, I didn't understand the question, and maybe Bart Proger could go back to it.

Proger: All right. Let's take only one type of sampling. Let's stick with one that Glenn is talking about. Let us consider a certain descriptive type of problem in arithmetic, and let's stick with the learning disabled kids who have perceptual hangups, again. If I sample randomly, the various idiosyncrasies should come out in the wash, but they don't in actuality. They are still there. When you say, I'm going to sample from total item universe in a particular area, you are going to pick a lot of those hangups up and your test won't really sample the content domain. Instead, you will be measuring perceptual impairment. I think Glenn is saying we can partial this out if we do two samplings and try to identify test items, both with respect to perceptual hangups and for straightforward, nonperceptual problems.

French: It seems to me that you get a universe of items, and these could be placed into the computer, and that you could sample these for straightforward testing if projects miss some. Then you go into a branching program which looks at all of the tests, and then if you are successful in the sampling, you go on to some other concept.

Millman: Are you asking about the case where a child's quite low score on a test might be invalid because, on the one hand, the items he could do weren't included in the item sampling, and on the other hand, the items he did poorly on embodied perceptual peculiarities?

Mann: I think he was asking that.

Millman: If you are concerned about that...and have to get diagnostic information about every possible thing...I don't see anything to do except complete sampling.

Proger: I think what Glenn suggests—sort of stratefied sampling pre-identifying certain known areas of difficulty, and adding these to the usual test item domain, that this might pull out the things that a diagnostician would want to identify. If you just sample regularly, the way we would for normals, and say, you know, I'll just take a collection representative of the total universe of items, it would fail. I think Glenn answered the question very nicely—that if you pre-identify those problems that you know the kid will have hangups in—you know, visual perception, reversing the numbers and so on—that you can say, all right, with these problems we know if he's going to do poorly, but the others he may very well have complete mastery of, and we might be able to say that the mastery of that arithmetic operation is intact for the kid, but that a visual perceptual problem may interfere with assessment of that operation.

Mann: A major problem is the variability of the population, as well as sampling. You may have a population of educables, so to speak, and you can have an item universe established for them. However, within this group of educables you have youngsters ranging very widely in ability. How does this affect our sampling?

Proger: Even "EMR's," let's say, of moderate ability range, may, for sampling, constitute a broad universe, but it is not like normals. We don't have a normal curve.

Millman: The purpose of the testing, though, is not to just make judgments about the individuals—is that not right? Then the fact that the various people you are going to make judgments about are extremely variable themselves, doesn't negate the use of standard test procedures. For individual assessment, you may want to proceed along criterion-referenced measurement lines. I don't see how the nature of the groups (gifted, retarded, or whatever) mucks up the water.

Mann: Special education implies special population, and individualized or prescriptive instruction, in one form or another. Prescription is a very popular term now. It has an air of prescription to it and provides specific improvements. When Peter's book came out on this subject, he offered rather gross prescriptions. Go see the county psychiatrist or your guidance counselor. But today, in special education, we see people saying that we can really identify certain areas of dysfunction and then provide a program to correct them. Now we already discussed the pros and cons to this kind of approach, but let's approach it again, in terms of are there ways that we can validate educational prescriptions, if you will, with techniques of remediation, intervention, compensation, training, etc., and can this be done so as to finally accrete a corpus of knowledge that would allow us to identify current prescriptions for individual cases, and can we really validate curriculum in this sense?

Proger: Along these lines, we have precision teaching. Kunzelman and some boys out from some university in the midwest asked the teacher to watch how the kid performed on, let's say, a set of arithmetic problems—giving a bunch of worksheets or that, the usual day-to-day activities—and they say you can really see how your individualized prescriptive package is working for the kid by looking at the rate: number of problems correct per time spent. So they say, don't look at absolute levels of performance, but look at the rate of mastery. What problems do you see in this?

Mann: In short, is there some way we can systematically and intelligently validate a body of interventions for handicapped or even with normal youngsters?

French: Well, the problem as I see it is, what kind of instruments do you use to determine the prescription? If you try to use a Binet or WISC to determine a prescription, you are without any help at all, or at least if there is some help, it is through knowing that they are severely retarded or moderately retarded intellectually average. The question, I guess, for me is, how you do analyze anyone's performance to determine what the teacher ought to do next? It seems to me that it gets it down to very sub-partials—you know, almost on a day-to-day basis. It is very difficult to develop an individualized curriculum a year in advance. Maybe one could develop a course of action for several weeks, but then we would need to re-assess progress and determine whether or not the system has been effective.

Feldhusen: The precision teaching people are a branch and really not much differentiated from the behavior modification movement, and so first off, their own views of what is necessary to validate or support the efficacy of procedures and so forth, are totally different from the kinds of experimental designs and approaches that we would have in traditional education. Ogden Lindsey uses that term too, although his brand is something else. He was one of the first people to speak of precision teaching. But again, the proof of the pudding in precision-teaching for all of its advocates, and here is where we fly head-on into two different positions—their proof of the pudding is purely in what they can show for one child. End of story. If you can do it for one child, there is no interest on their part for the group, long-term effects and so on. There is, in fact, much arguing by them against the obfuscating effects of evidence about how you did it for ten, because then the water is all muddied up with some people who didn't make it and what have you. But, I believe somehow or other, you have to come up with a mean. I think my assertion is quite safe that here we have a head-on clash of two stances with reference to evaluation or researching something.

Millman: From what I understand, John and I would appreciate a little more amplification. Specifically, if the behavioral modification of precision-teaching type of researcher was asked to show that he was effective with the ten youngsters that he has been working with, how would he show that he was effective in teaching them how to add two-digit numbers?

Feldhusen: Of course, he could cop-out like so many people in the Journal of Applied Behavior Analysis. They tell us about one dramatic case—you, the Skinnerian position—is that a large dramatic, powerful effect disdains significance-testing for a threshold of significance, or what have you. In effect, the result is so big that it hits you between the eyes. You don't need to worry about significance tests, and so forth.

Millman: Would precision-teaching do something like, say all right, here is Tommy Smith, and before I started prescription-teaching or behavioral modification, little Tommy Smith didn't know how to carry numbers in addition?

Feldhusen: He will state the rate now. He was very precise, and the word "precision" is involved here, because he is going to be using a counter for sure, or some way of enumerating the incidence or frequency of behavior.

Stanley: Skinner would say: "It must work on every rat or Skinner is not interested." I heard a Skinnerian student—a student of his—give a paper in which he gave four rats and apologized for the extra three on the grounds that he just had them available. This is the extreme case, but he said it must work for every rat. He is not interested in something working with two rats out of four. But, it does seem that too often, we get hung up on the control/experimental group thing. I am guilty, Campbell and I certainly, of promoting this point of view in extreme, almost. But an illustration somebody may have heard: "Years ago, I had ridged, infected fingernails from a fungus, and had had it for some twenty-five years. They were troublesome, but they didn't hurt or anything; they didn't look very pleasant. And, I went to dermatologist after dermatologist, and he removed them, pulled them out by the roots, treated them with iodine, did all kinds of things...nothing did any good. Twenty-five years. In 1959, I ran across some little, obscure note that I had made in Science, that some fellow over in Germany developed something that worked on fungus on rats. So I rushed down to the dermatologist, at the University of Wisconsin—the Professor of Dermatology—and he said, "Oh, yes. Got an experiment going on—I'm going to get forty people and see whether I can cure fingernails by oral ingestion with the stuff." Well, four months later I had good fingernails. It was not really necessary to have any control group. The alternative hypothesis that my fingernails could have cleared up spontaneously, concurrently with the administration of the stuff was so improbable that even if it hadn't worked with the other thirty-nine, I would have been relatively sure it had cured my fingernails. And, I think, I could have made a strong case logically for that, in the absence of other things. Now, it did happen that all forty of us had gotten our fingers back in six months. We all had good fingernails, and still have good fingernails thirteen years later. No control group was necessary. It was a very sloppy study, as studies go. Sometimes, they took samples of this, sometimes that. The poor professor couldn't get his interns and residents to do the thing right, so he didn't get decent data to analyze, but he showed conclusively that the stuff cured that type of fingernail involvement. Now, control group of two or three subjects would have been powerfully impressive. One subject was powerfully impressive, four or five would have been a demonstration of a minor medical miracle. And, forty was just so overwhelmingly convincing—a huge effect.

Mann: If I brought you a youngster, though, who had a reading problem, and I had applied so and so and got dramatic results, and said this proved that my technique worked, you'd say, "Come on now, this is a trifle naive."

Stanley: Well, see, we wouldn't know what you did. We know I just took the stuff—that was the only thing.

Mann: You would know what I did. I used a reading kit.

Stanley: The fact that I happened to take up jogging at that time, you see, was an implausible alternative explanation—if I had done that.

Mann: Now in medicine, we accept this kind of demonstration as being legitimate proof in education.

Stanley: Well, it's a Skinnerian-type proof, though. You've got the common graph doing this, and suddenly you change something and you say, aha. I predict that when I change, the rat is going to do this, and if he doesn't do that, then I don't have a variable I am interested in. I have to find a variable for every rat that, at some point, I introduce this different variable ratio, and he moves in a predicated direction from then on.

Feldhusen: Let me support, by a different illustration, Julian's point of view. You seem to suggesting, Les, it is fine and dandy in the medical world.

Mann: No, what I am saying is that nobody would accept a single case as proof of the success of intervention in educational problems.

Feldhusen: Not so. You are missing another part of the word; well...no you are not. Let me try to explain. Another part of that word, precision—and this whole thing turns me on a little bit—the other part of the precision-teaching asserts that the problem with the educational treatment you find so hard to make an inference about is, that it is often such a gross undifferentiated bag. So, what always characterizes precision-teaching is precision in definition of the behavior to be changed. Now, that often means that you come down to something as simple as a pidgeon packing a disk, or a child doing a particular kind of problem. A very specific behavior objective is always implied in precision-teaching. Furthermore, your treatment is always going to be very, very specific. That precision, I think, implies, and I think, does permit, when this kind of a thing happens, does permit relatively clear inferencing about the relationship between treatment and effect, as compared to so many of our treatments in education which have been so large, gross, and undifferentiated.

Millman: The difference between Julian's treatment and Les' bit, it seems to me, isn't that we don't know what was in the kit. The difference is in the fact that there is no alternative that's plausible—not alternative explanation, but alternative treatment—plausible for Julian. I am afraid, however, that I am missing the point about precision evaluation. Let us take a case, because I think it is easier to work with special cases. A student is, right now, confusing lower case "b" and "d," and you are trying to get him to make that discrimination. That is a very precise kind of thing. Now, here comes Johnny Jones who is all screwed up. He can't make that decision, or that discrimination, at this time. You knew you could modify his behavior. You have some sort of training technique, and then you say, look, in thirty minutes of instruction, I got where he could discriminate thirteen "d's" and "b's" and every one correctly. Is that the way it would go?

Feldhusen: Yes, that is the way it would go.

Millman: Well, in education, I don't think the argument would be that it was the treatment that had the effect; I think the argument would go...yes, but there are other ways of getting him to discriminate "b's" and "d's" so what is so red hot about the precision; why should we move over into your precision training? Why don't we just use this other approach. I mean that one could argue that a worthwhile goal is another matter. There are other ways of reaching it. In Julian's case, there was no other medicine.

Stanley: There were. Other methods were tried.

Millman: But in my case, no one would argue that the training wasn't responsible. No one would argue that the treatment didn't cause all these children, out of nowhere, to be able to make these discriminations. I mean, I certainly would agree that if the behavior modification thing worked, but there are other things that could also have worked. There are other ways of teaching kids. Do the precision people say, well, I don't care about these other ways. My way works, and I don't care about others? Is that what they would say? I don't care, and it is true it took me thirty minutes, but I don't think anybody could do it quicker than thirty minutes.

Hagin: I would want to know one other thing. When this child reads a book, will he also discriminate "b's" and "d's" then? I will propose another experiment. When our engineered classroom went into effect, it was felt that if the child attended to the reading task, his reading would improve. And the children did improve in their ability to attend to reading. I think the child read for three minutes consecutively and so on, but when the children were tested on reading at the end of the experiment, the results were very unimpressive. This same technique with attending to math worked pretty well, and they did manage to change math scores, but the nature of the thing that was being taught determined whether the technique affected the instruction.

Millman: If I were on the behavior modification bandwagon, I would say, well, you just switched the rules of the game on me. You are interested in behavior over here, when in fact, I thought you were interested in this one over here. So, that is what I modify. So, if you want me to have kids discriminate "b's" and "d's" when there is context, then okay. Now, if that is what you want, then I'll go ahead and do that. But, don't keep switching on me.

Hagin: It seems to me to be that generalization is always necessary.

Stanley: May I mention something we are doing with very bright mathematically and scientifically talented youths? And, I think I will illustrate some of the complexities of the problem. A few years ago, we started out, just incidentally, with this study...this was before I had this big grant. We discovered a very precocious boy in our computer center at age twelve in seventh grade, teaching our people FORTRAN—teaching graduate students how to do FORTRAN that came in. He was unobtrusively helping them orient their data. Well, he came to my attention, and by the middle of eighth grade, when he was age thirteen, I had given him some college board tests, and he scored in the middle to high 700's on the math and science ones...a very, very talented boy. So we worried and speculated and finally we decided that since, clearly, he was not learning much in public school, we would tentatively introduce him to Hopkins as a regular student, and let him take some courses he looked particularly good in, and if he did well, we would just let him stay, and let him be a regular student. We would not let him go to high school: ninth, tenth, eleventh, twelfth grade, at all. Well, this was a very radical proposal and his father speculated for nearly six months about it or nine months before he agreed. So, we brought him in and let him take advanced—take honors calculus, and sophmore general physics which has a calculus prerequisite; he had already made 772 on the math II of the college board. He was pretty well oriented in math already. And computer science. Well, he made a 3.69 GP on a 4.0 scale. He made a high B+ on calculus, and a solid A in physics, and a very high A in computer science. Now, okay. The next year, lo and behold, if somebody doesn't hear about him through a very obscure mechanism, and a mother comes pounding on my door, saying, I want my son in. He is just miserable in eighth grade. Boy boys throw rocks at him and they hate him, and they could kill him if they could. Now, he is just utterly miserable. So, we had him take college board tests, and he scored in the middle 700's on several of them—math and physics—and even scored well on chemistry. He had been fooling around in his father's junior college, and looking over peoples' shoulders and things like that. Now, here was a boy who looked neurotic, maladjusted, upset, socially and poorly adjusted, you know. Ordinary prescription would be, make him get adjusted where he is before we let him do anything else. But no, we finally let him come to Hopkins, and he has been even more successful than the first one. He is fantastic. He's unbelievably successful. He is a fifteen year old sophomore now, taking a 50% overload and making practically all A's in all the hard courses. Just doing enormous things. He's going to make Phi Beta Kappa without any trouble. He's making all A's and B's, and he takes nothing but advanced hard courses—math and philosophy courses. Okay. It worked beautifully. It is implausible to suppose, in either case, that somehow the boys spontaneously began to to get along a lot better. It is quite possible to suppose that the introduction to much more stimulating environment was the cause. But, now suppose I have another boy who looks to me, for all intents and purposes, somewhat like these boys who are personalitywise very different, but academically and aptitudewise looking like identical twins on paper. Do I have any assurance that the next boy will work? There are so many factors in the backgrounds of these boys, that I am still very cautious and reluctant. I still don't know the dimensions of what we do in order to put someone from eighth grade into college—or from ninth grade into college—and have any assurance. It is not like my fingernails. I am pretty sure that my drug will cure the next person, and the next one, and the next one, if he has the particular fungus involved; there is a relationship, but I am not sure about these boys because there are so many chance things in the environment...I don't know the moderating variables. So, although we will keep trying, we're working with human beings, and we will never be able to control the environment well enough to be really sure for all our cases.

Feldhusen: Bayes has been doing similar inferencing, which was pretty dismal, with reference to something quite seemingly, quite similar, to what you are trying to do. So, that one of these are probabilities of failure.

Stanley: These are commuting students, for one thing. These students, a lot of them are residential, and so forth. Also, these are quantitative math folks, and the Chicago students were heavily verbal. My guess is that precociousness is much less in the verbal area.

Mann: This gives us a chance now to re-raise a question. We do have a fantasy in special education that by trying out various techniques, with certain kinds of youngsters, handicapped particularly, we can somehow validate these, as far as effectiveness goes, and put such information into some massive data bank, and then eventually—sometime in an undetermined future—press buttons, identify with pupil as best we can, according to the characteristics...the problem he has and get the best fit prescription, if you will, of intervention for him. Is this practical? As you said here, you had one student come in and this works while another one with similar characteristics nothing will happen, perhaps.

Millman: But, what you are looking for is the aptitude-treatment interaction; aptitude in a very general sense. The underlying assumption effective is that there exists these interactions. That is, when a kid is given this kind of prescription there is something that is going to be differentially effective with him. Well, we have discussed that earlier, and we had pessimism around this table regarding that. I may be imposing too much of my own opinion, but unless you look very hard at what the learning task is, and perhaps prescribe very close to that, you won't get anywhere. On generalized tasks nothing much is likely to result in the way of useful prescriptions.

Bracht: Where your approach may be useful, Les, is where we have a curriculum which consists of hierarchically organized tasks, and given a certain task that we want to teach, we can use information from performance lower in the hierarchy to see what are the prerequisites for this task. We can look back in the student's record and see if he has the skills for these tasks. If he has, we can proceed with teaching that. If not, we can go back and teach those tasks which are prerequisite. It will require that we will then have a comprehensive task analysis of the curriculum for the purpose and criterion-referenced tests for each of these tasks.

Mann: This is not necessarily aptitude-treatment interactions that we are talking about here.

Bracht: That would be one variation of it.

Feldhusen: I think this echos my earlier assertion that we have sort of a rough choice of sticking to characteristics, and I wish that aptitude-treatment interaction had come out CTI, characteristics-treatment interaction—a choice of looking to things that relate quite specifically to the school learning problems, or a choice of looking to things like elements of the structure of intellect model or personality defects or perceptual disorders or what have you, and it seems to me, for the present, and Glenn has said this—I am just echoing it—the most hopeful place to look is to the child's background in learning tasks, as related to the new task that we are trying to teach him. Or where we detect a deficiency. This, I see as much more hopeful for the present than looking to elements of structure of intellect model or personality deficiency, or what have you.

French: Well, yes. I am more optimistic than Jay. And along the lines that Glenn was saying, I think that the instruments that we have for evaluating student performance are not very good. But, I hope as we move to a more effective means of keeping track of the individual objectives that are obtained in each of the wide variety of subject matter areas, that we will be able to plan which of the—maybe an infinite variety of things—may be best done at a given point—what the next learning task will be.

Millman: There is another kind of data bank you could have, where you put in various problems, or various situations, and then put in the results of all different attempts to deal with it, and then you begin to accumulate an experience about what will work. One problem might be, I am really searching, the problem might be, that the kid won't put his rubbers on when he goes out when it is wet outside. The kid won't wear his boots, you know. There might be twenty different ways of handling that situation. You might be able to build up this bank of experience...different studies...but that is quite different than talking about aptitude-treatment matches.

Hagin: I'd like to see this bank that you fantasize, in terms of content and possibilities. But, I would also not like to remove the facilitator—the teacher—the person—who will take a risk on an emotionally disturbed youngster, putting him into college or... Having seen the life history of some of the kids we worked with, I don't know that this bank can ever fully be adequate for individual needs. I don't think it would ever reach the complexity where the risk-taking in decision-making can really be reduced significantly.

Mann: I would agree.

Hagin: Done arithmetically, so that the human element of the teacher-facilitator, the person, who can take into account strengths, weaknesses, motivation, and so on.

Mann: Ken?

Hopkins: Matching some treatment with a particular type of child, of course, in a simplistic and degenerate sense is nothing new. We don't put children in school till they are six years of age, and we have a curriculum for that kind of youngster. We don't put people in Algebra II before they have Algebra I. That is the same idea in a very crude molar degenerate sense. So in spirit, I favor that, although I am not optimistic about how far you can go along those lines. Particularly if the teacher remains a component in the instruction. In theory you know, if you can measure all of these things, with, of course, enough research, you can find the optimal fit. But, some of the critical variables, I am not at all very optimistic about, being able to measure them both in the teacher, and on the part of the student. I think we can measure the cognitive aspects pretty well. But the other factors—motivation and all that, are going to be more difficult, and I am not at all sure that we won't find ourselves in a trade-off situation if you try to control these; you know, if you take the teacher out of the situation and make it completely computer-oriented. I think you have a much greater chance of fulfilling the ideal of a controlled situation, but you may lose a lot of the cognitive input produced by the teacher and the gains that go with these.

Mann: Computer instruction, you are saying. Direct computer instruction?

Hopkins: That is under your control. You can make sure of the matching.

Feldhusen: Jack Sawyer has, in a sense, provided final answer to that problem. I heard Rosa saying, seemingly, that she would not want to remove out of the teacher's diagnostic functions—looking at and deciding what is really wrong with the child. ..what he needs. Is it fair to say that you were saying something like that? Jack Sawyer, in a very interesting review of the old problem—computer versus human judgment—finally came up with the ideal solution. He said that that is not the right way to conceive the problem. His resolution is: yes, from the literature, you must conclude that the computer is going to do the best synthesis of the available and make the best judgments. It will beat the human being. But, the ideal solution is to use any and all objective data, precise tests or what have you, but teacher judgments, or clinician judgments, or whatever they may be, can add significant increments in the analysis or the prediction. So, the final judgment here, is to let the computer be the final decider, but to let the teacher or clinician, or whoever it is, make a substantial entry or judgment because it does seem, in quite a few instances, that human judgment can provide kinds of insights that go far, at least at present, beyond what we can get off from objective tests.

Millman: Dick Barrington, whom I greatly respect, says that the

Sawyer article is not the final word, and he is quite critical of that article.

Feldhusen: Alas, Jack has dropped the matter, too, so I hope somebody picks it up and continues the battle.

French: I certainly wouldn't want to be in the camp that says that we ought to computerize all instruction, even though we ourselves are working in that direction.

Mann: Well, computer-assisted is different from direct computer-instruction.

French: Yes, I think we need to try every possible way to try to help the teacher make those important decisions. I like to view it a little bit like going to the airline ticket counter, and have them push a few buttons, and they come out very quickly with a reservation or a possiblity of going somewhere. If we just had some way of helping the teacher put information in and improving her memory through the use of the storage system as well, I think it would be helpful.

Stanley: You know, I am reminded very much of one of my favorite quotations from Percy Bridgeman. Something to the effect that the scientific method insofar as it is a method, is nothing but doing one's damndest with one's mind...no holds barred. And when I think about the question of whether we should look at aptitude or characteristics of the task, well obviously, the answers are nonsense to propose it that way. You look at both—the interaction between both—and you are concerned back and forth. You are wrestling with it as a wrestler might. He doesn't decide in advance to use only a headlock; he'd lose most of his matches. He has got to adapt to the situation. One we overlooked, is the fact that no matter how we try, there is going to be a lot of slippage in human beings. About one-fourth of the Hopkins entrance, no matter how carefully we select them, don't make even a 2.00 GPA the first semester. A fourth of them are unsatisfactory. Some of them do much better than we predict. Some much worse. There is a lot of slippage. Now, an illustration of how we might look at both of these two boys I mentioned that were so precocious. Both took in the first semester of the sophomore year...they both took symbolic logic...and that sounded right for them. They are real hot-shots at computers and math and tic-tac-toe, and so forth. But one of them made a "C" in symbolic logic, and the other one made a "B," and the one that made a "B" is going to be a mathematical logician, and he was concerned. He wanted to know why in the world he couldn't make an "A" in symbolic logic. He made six "A's" and one "B," and that "B" was in symbolic logic. Well, we began to think. What are they? We weren't thinking about the didactic aspects. We are not emphasizing the teaching—the didacticism. We are emphasizing the learning. Why did he fail to live up to an "A" level? The other fellow even up to a "B" level? And we looked back and we surmised, and we talked it over with them—I am using all kinds of clinical developmental methods—we talked with them and decided that probably they were simply outclassed verbally by their classmates. They were people operating at about, you know, 600 GRE verbal, which is awfully good for a fourteen-year-old, but not very good for a graduate student at Hopkins, or even for a philosophy major who is a junior or senior. So, they were clearly outclassed verbally. They needed a 700 verbal, at least, to make an "A" with appreciable effort...maybe even more. So there we had an illustration of looking at an antecedent measure, or verbal, and looking at the task and interrogating the person being taught, back and forth, until we have a tentative hypothesis. That we have got to have more verbal ability on the part of these youngsters before they take symbolic logic, because no matter how well they solve the three dimensional tic-tac-toe problem, they can't communicate to the teaching assistant, or the professor, an "A" level of solution. They were communicating on the "C" level. So, I think in that sense, we can clearly look from the aptitude measures up to the task and back again.

Mann: Are there any other contributions, as far as the idea of a national data bank of prescriptions or instructional techniques for instructional objectives? Julian, is it worth conceiving, or having a data bank, where various techniques that have been tried out and have been found successful in varying degrees, are fed into a central place where you can have a computer printout to assist the teacher in managing the pupil?

Stanley: Yes, this thing might be helpful. But, the real problem is the—what we have been discussing—complexity of learning, and the fact that learning is conditioned on other factors. But, I think any kind of rules of thumb that can help the teacher are helpful, so long as they rest on a reasonably sound empirical base. That may be the problem...so long as they can be reasonably qualified. Now, how do you do this? How do you communicate this to a teacher? First of all, that we really don't have enough panaceas—things that really work. It is difficult to generalize about findings, in that they are robust with respect to individual variations in circumstances, didactics, and pupils. But, insofar as you can communicate something to the teacher about what is known, and program it for them, and keep them from being too stereotyped—okay. However, with really hard-to-teach youngsters, training, rather than education, is the mode. I would call it training, some carefully sequenced programs, with brighter and able students. It seems to me that the emphasis ought not to be on didactics, but on learning...on the youngsters' learning to use their various talents and facilities to get the material, rather than being taught the material.

There is a difference between what I call training, and education. Training is important for educable mental retardates, for substandard naval recruits at San Diego, and things like that. There is no other way to get out of it. You cannot lecture to them, and have it back. You can't say, just go off and learn to use the computer. But at some other level, average and above, I think there should be less and less emphasis on sheer didactics, and more and more emphasis on pupil-motivation and behavior with the materials available.

Mann: Piaget, and others, have suggested that many subjects should not be taught...all, probably—but give the youngster books on the subject, and answer their questions.

Stanley: Well, that this simple thing like Albegra for bright—the kind of kids who take Algebra, you know--the somewhat above-average youngsters, in a typical junior high, eighth grade. We have two beautiful examples—two contrasting examples in Baltimore that I noted. One is a fellow who somehow in eighth grade, junior high, managed to put most of the kids on self-instruction, and by the end of the year, a lot of them already knew two years of Algebra. They are pretty bright kids, but they learned it. Now those youngsters on our math aptitude test, the SCAT level IC will earn scores practically perfect by the end of that year. They'll earn 49 or 50 on a fifty item test. Equally competent youngsters verbally, equally promising, who have had the standard Algebra treatment, will earn scores of more like 35 out of 50. Now, ascribe most of the difference to the fact that the man had got these people to be self-learners. They have covered more material better, largely on their own. He is not a teacher in the didactic sense, not so much as most Algebra teachers are. He is a superior teacher. He has learned how to get people through two years of Algebra in one year.

Mann: Managing is a critical teaching skill.

Stanley: Right, not the sequence and didacting. I think a training approach to these youngsters would be disasterous...well, not disasterous...it would be uneffective. They would learn only half as much in a year as they learned in less than half. It is well known, for instance, that those of us who have taught Algebra in high school know you can take a bright pupil and kind of individual-study him, and you know tutor him through a whole year of Algebra in about a month or two. So, in effect, the youngster can learn five times as fast if he's apt as he does in class when he is lectured to. So, my concern would be with any kind of data bank or other prescription if it encouraged excessive didacticism on the part of the teacher for all but the really handicapped youngsters. We need for the others, teachers who are master facilitators. He puts importance on individual study. He doesn't let them get off the tract and he monitors them so that he is sure that they learn.

Bracht: I wish I had brought with me something I was just reminded of, a very brief outline by Lee Cronbach, in which he outlines two conceptions of learning and evaluation, something of this sort. One is the form of instruction, which is of the didactic sort, where you have specific outcomes which you want to achieve, and we teach to achieve these. The other is, where we kind of provide an environment for learning, and we kind of provide the stimulant to the student, but the student is expected to learn things beyond what is taught in the classroom. I think this is typical of many things we teach in university, and especially in graduate courses, where the student is expected to learn things that are not presented in class, or at least, to go beyond the level presented in class, and in most cases we would expect some transfer also in the outcomes. The student would transfer to other areas of what specific things he learned. This has implications for evaluation as he points out. If we are doing a research study in which we are trying to bring about certain specific outcomes, I think we ought to test for those specific outcomes. My preference would be to use criterion-referenced measurements. If, on the other hand, we also expect learnings of a different nature; other things to be learned than the specific that we try to train for in the instruction, and its transfer and such. So, we need to test for other things beyond the specific skills we have provided in our instruction, and I think so often, we try to use one dependent variable in our research studies where we ought to have multiple measures. I think we tend to pour all these into one test, getting a global test where trying to measure outcomes on specific skill "A", specific skill "B", some transfer to other settings, and application in other settings. We ought to keep all these measures separate. We ought to get such scores separately, rather than trying to get one global score.

Mann: This brings us to a small point. How much time shall we devote to evaluation?

Bracht: Well, in my own experience in teaching is that, the first time I am teaching a course, the tests that I use are not what I consider very good, and the next time they'll be a little bit better, but I think with renewed experiences, I can start to approximate what might be in my mind as an ideal goal. It takes a lot of work, I think, when we are first developing something; we probably need to spend more time on developing the instructional program and probably lesser emphasis is put on evaluation. Whereas, at later stages, we probably put more emphasis on evaluation. It is just that there is limitation on time. We can't do everything immediately.

Hagin: Sometimes, I fantasize this school I would like to have children go to, and I think it is about the opposite of what is done. I would like to see children start early, and to be taught tools as quickly as possible, so that the early grades would be much more formal and didactic, and then as the children are free, they would have much more opportunity to use their tools to find what it is they would like to know—in the realm of knowledge. Now, some of the experiments that I hear about, of freeing these children before these skills are completely learned, I have real questions about these, because I know how many of the children do not get the skills and skim the surface, pretending to sample realms of knowledge that they can't handle, in terms of basic reading and organizational skills. So, I'd like to see a kind of freeing as the children progress in the higher grades. As far as the data bank question of teaching, or a replacement to the teacher, but as I wander about in classrooms, I am impressed with the limited alternatives the teacher is offered. And there are as many things that could free the teacher to do the personal interaction with the children would seem to be the tool, as things now stand. Certainly, then, the data bank is not to be discouraged, but to be encouraged.

Stanley: It seems to me since we are getting back to the data question —it seems to me that we have forgotten about the diagnostic test. You know, we have been so busy with aptitude tests and achievement measures, that we have forgotten about diagnostic tests, but one of the things, for instance, we find a bright, mathematically precocious sixth grade boy, say. One of the things we do early is to give him a twelfth grade arithmetic fundamentals test. We give him both forms—we try to see what it is he doesn't know. Our test isn't a particularly good diagnostic test probably, but if he misses eight items out of sixty, we see what kind of items. Well, they are concerned with exponents, so my associate, who has been a junior high math teacher, teaches him about exponents for an hour or two, you know, then gives him the other form, and if he makes a virtually perfect score, we go back to the school and say, let him out of arithmetic. He no longer needs arithmetic; he is good at reasoning, too. Now, we will put him on a special math program on individual study, and let him work with his teacher on a tutorial arrangement. Let him learn it on his own. I think so often, we grind these poor kids through and through and through and through things they already know 90% or 95%. We have eighth graders in general science, college level. Now, those kids, in effect, know more general science than many college graduates. They really know more general science, and yet they are forced to take eighth grade science day-by-day in some kind of didactic or individual study, or whatever sense, and they don't like it. They go all the way from day-dreaming to resenting it, you know. Now, I think that kind of thing would free the teacher from lots of futile evaluation of little specific petty aims for some of these youngsters, and let them get on to physics, chemistry, for the general person, and some more advanced math for the math person.

## REFERENCES

Armstrong, J. R. A model for materials development and evaluation. Exceptional Children, 1971, 38, 327-334.

Baker, E. L. The technology of instructional development. In R. N. W. Travers (Ed.) Second handbook of research on teaching. Chicago: Rand McNally, 1973. pp. 245-285.

Berlinger, D. C. & Cohen, L. S. Trait-treatment interaction and learning. In F. N. Kerlinger (Ed.) Review of Research in Education: 1. Itasca, Ill.: F. E. Peacock, 1973, pp. 58-94.

Bogatz, B. E. An investigation of teacher experiences of instructional materials. Exceptional Children, 1971, 38, 233-236.

Bracht, G. H. Experimental factors related to aptitude-treatment interactions. Review of Educational Research, 1970, 40, 627-645.

Brown, L. T. A behavioral index of the exploratory values of prose materials. Journal of Educational Psychology, 1972, 63, 437-445.

Coleman, E. G. Collecting a data base for a reading technology. Journal of Educational Psychology Monograph, 1976, 61, No. 4, Pt. 2.

Fassler, J., & Brant, N. D. Teacher observations on using ear protectors for disturbed children. Exceptional Children, 1971, 38, 254-256.

Lumsdaine, A. A. Instruments and media of instruction. In N. L. Gage (Ed.) Handbook of research on teaching. Chicago: Rand McNally, 1963, pp. 583-682.

Mitchell, Jr. J. V. Education's challenge to psychology: The prediction of behavior from person-environment interactions. Review of Educational Research, 1969, 39, 695-721.

Popham, W. J. Curriculum materials. Review of Educational Research, 1969, 39, 319-338.

Proger, B. B., Carfioli, J. C., & Kalapos, R. L. A neglected area of accountability: The failure of present efforts in instructional materials evaluation and a rational solution. Journal of Special Education, 1973, in press.

Salomon, G. Heuristic models for the generation of aptitude-treatment interaction hypotheses. Review of Educational Research, 1972, 42, 327-343.

Sanders, J. R., & Cunningham, D. J. A structure for formative evaluation in product development. Review of Educational Research, 1973, 43, 217-236.

Smith, H. A. Curriculum development and instructional materials. Review of Educational Research, 1969, 4, 397-413.

Ysseldyke, J. E. Diagnostic-prescriptive teaching: The search for aptitude-treatment interactions. In L. Mann & D. A. Sabatino (Eds.) The first review of special education. Philadelphia: JSE Press, 1973. Vol. 1, pp. 5-32.

## CHAPTER 6

## DIAGNOSTIC-PRESCRIPTIVE TEACHING AND THE GUILFORD STRUCTURE-OF-INTELLECT MODEL

Overview: The present chapter provides additional insights into educational prescriptions by means of the Structure-of-Intellect Model (SOI). Guilford has postulated his SOI model of cognitive thinking in terms of a massive, three-dimensional cube divided into 120 cells. Each cell represents a highly specific level and type of mental functioning. An implication of the SOI model is that tests can be found, or built, to diagnose each cell and that then a prescription can be built to remediate, train, or teach to that particular ability. However, as the participants point out, there are many difficulties confronting the model. For one thing, no one has ever been able to analyze samples of on-going, real life, academic work into all of the 120 skills of SOI.

Mann: Shall we consult the Guilford model, and factor analysis as a basis for corrective design? Guilford's work is mentioned repeatedly as a rationale or justification support for various positions in special education, i.e., diagnostic-prescriptive interventions, Joe.

French: Well, I guess the importance of his work is his causing investigators to look for minor, or subparts, of the intellectual process as it relates to a particular academic activity, i.e., trying to force investigators away from global measures into more specific ones that can perhaps be more isolated to stand on their own. He feels that after divorcing individual measures as much as possible from other individual measures, their composite gives us a better picture of the individual than some more heterogeneous device.

Feldhusen: I am equally ignorant about this so-called prescriptive teaching. I have a vague notion about what that is all about. Are you saying that the many tests that emerge out of the little cells of the model, are used as a basis, first of all presumably for the description of the child, and then for the prescription of instructional treatments that somehow or other are inferred from one or another element of the Guilford tests?

Mann: There's the prescription that his tests isolate ability and that you can train, or teach, to these—even remediate these. Now, Meeker had a curriculum for gifted at Southern California, based on the Guilford model.

Stanley: Who has demonstrated that they are important? Well, in 1959, I think, or 1960, Guilford made some educationally prescriptive statements based on his model, but he gave no validation of those statements at all. He merely said, you know, using Descartes, paraphrasing Descartes, these abilities exist, therefore, they ought to be taught, sort of. Now, I wondered, have people done task analyses? Now there are two different points of view—one point would be, we have some kind of school task which requires certain abilities or entry skills. The youngsters might have to have these developed, in order to do the school task. If they don't have them, then you train them accordingly. Another approach would be that there exists certain abilities that are important and should be developed because probably in some school task later on, they may well be very useful. I would be much more suspicious of the latter model. I am not sure: (a) abilities in a generic sense are trainable, and (b) I am not quite sure that if they can be trained, that they will improve performance in such a way that the ability naturally seems to covary with the performance. But what is the state of this whole thing?

Proger: It is assumed that one can do prescriptive teaching on the basis of the weaknesses and strengths that one might find testing on the basis of the model. There, you assess certain abilities and intervene, or help deal with, some of the problems the kid exhibits in the total model.

Stanley: He might not be very good at certain things in the model. That doesn't mean that he is not very good at certain school enterprises. The test-training model is one that Bereiter and Weikart, particularly Weikart, used for preschool facilitation. The results, you know, have been very disappointing because when you raise IQs artificially by force-feeding, they lose their predictive value for school success. The natural IQs correlate with school success, but raised IQs of the Weikart sort tend to correlate less well. This is a general finding in psychology. The predictive value of, say the beginning test when a person goes on a job, is greater than after he has been on the job for a little while; after he has been trained in the task. Fleishman found this, too. How does the model go, again, for special education?

Mann: That there are basic skills and subskills that are basic to learning, academic and otherwise, and these can be identified through testing, or other evaluation, along the model's lines. You can develop, or stimulate growth, in these abilities, train them or correct them, or compensate for them by utilizing others that are more salient or stronger or better developed. Doesn't Fleishman recommend this, too, that you train different ability structures at different ages?

Stanley: The Binet is the best example of one that has been talked about. Yet, I am not quite sure that anyone has demonstrated quite fully that simply raising Binet scores by playing with Binet's materials, in effect you know, improves learning ability. It just improves Binet ability. It doesn't transfer to Raven's Progressive Matrices, for example. So, I am skeptical of picking out an abstract trait model, and then studying in detail whether or not the improvement of the ability really follows. It is like this—it is a correlation of situations. Suppose we find a pretty strong correlation between high-jumping ability and running the thousand meters, or hundred yard dash, say. Now, we train a person intensively in high-jumping—we make him a better high-jumper. It does not follow that he is a better runner. There is some kind of natural correlation, some kind of physical fitness correlation between high-jumping and running—improves his running. Although naively, you would think that it would, but since correlation is not causation, you don't. So, it remains to be demonstrated the improved high-jumper does, in fact, run better, nor does improved running affect high-jumping better. It may be quite possible that neither happens.

Feldhusen: We must come back to the house on which this whole structure of the intellect would be built. I guess it is fair to look upon it as a house, and might it be a house of sand after all, and an unprofitable way to proceed in our research. There are powerfully strong people who have made devastating criticisms of the structure of intellect as a respresentation of intellectual functioning. First of all, quite a way back, the review by McNemar, which was not specifically focused on the structure-of-intellect, but on efforts to fractionate mental abilities, that was an article in the American Psychologist, in 1962 or 1963, which first impressed me that maybe this fractionation business would be so profitable. Recently, a document I have alluded to all too frequently, already the review by Cronbach and Snow, takes a sidetrack for a lengthy review of the structure-of-intellect model, and comes out with very devastating opinions—the kind of criticism that, more or less, says that this is an unprofitable result and not a worthwhile product. Then this is the extreme pessimism of Chester Harris with reference to Guilford methodology. So, there are such very powerful criticisms that I must ask, is it worthwhile proceeding to study it? The kind of methodology that Julian seems to be suggesting is that we should see if the structure is a tenable basis for assessing abilities and then go on to see whether these are useful in devising remedial strategies. But, I think that to be seriously concerned about whether this basic model itself, even though it has become so popular, is worthwhile pursuing at all.

Stanley: There is also Carl Bereiter's research. He has tried both ways; he claims he has tried raising IQs per se, and he has tried teaching reading skills, you know, the precursors of reading skills, per se. And he claims that he gets as good an improvement in IQ by concentrating on reading skills as by concentrating on IQs. In addition, he gets better improvement on reading skills, too. So, there is the further question of whether one is well advised to concentrate on traits to the exclusion of concentrating on school tasks or the preliminaries of school tasks.

Hagin: As a practitioner, I would certainly want to underline this point—that we...our instruments are very fragile as it is, and when we start manipulating them, we are destroying something which threatens, perhaps, our planning for children. There is a second point—and this is the business of teaching to the test, and much has been said in this direction. I see it, if one would accept on the Rohrshach, the fact that the visualizing color is an indicator of emotional sensitivity, therefore, it would follow, if we were going to do this training on specific traits, that if we taught children to see color on the Rohrshach, we would make them emotionally responsive.

Mann: That is a good point.

French: I don't disagree with the things that have been said, but I do think that one of the advantages of Guilford's position has been to point out the broad aspects of the intellectual behavior, and I think at times we have been too narrow in our interpretation of what comprises intellectual behavior—and by calling to our attention the broader characteristics, we then engage in training programs that perhaps will help one develop more skills for someone who is already competent in other ones. For example, to try to develop skills in synthesizing information, when it may not seem to be developing as rapidly as one would expect from other measures, or work on evaluative skills in someone who is competent already in divert recalls.

Stanley: A very important point, because in working with youngsters of any kind; in working with these very bright youngsters, we've discovered that this—well, we didn't discover this, it is well known, we just reconfirmed it—you take two eighth grade boys, each with an IQ of 160, you know, Binet IQ. They may be very different in their cognitive abilities—they would be quite different. One may not know very much science; the other may know a lot of science. One may be quite poor in mechanical comprehension on the Binet kind of test item; another may be good. In other words, it almost appears as if some youngsters are very bright but not talented, and conversely, some are very talented but not very bright. So, you miss some youngsters with an IQ test, and you screen others in who are disappointing because they really don't have these cognitive abilities that are needed to be genuinely successful. So, I'm somewhat disagreeing with McNemar's position, in his presidential address that he made, and I think that that kind of thing is important—I'm mot sure though that Guilford's way is the way to go. There are many other types of tests available for such purposes. The question is, which ones are most relevant for what we are trying to do.

Feldhusen: I think that there is no disagreement with your position, because you characteristically limit your conversation to verbal and quantitative differentiations which is a very parsimonious differentiation compared to the 120 cells that Guilford proposes.

Stanley: I'm thinking of Raven's Progressive Matrices and tests of spatial relations and so on. There are a lot of full-length tests that are already commercially available. We are using them with our especially-gifted. Furthermore, we are using personality measure, Eysenck's Personality Inventory and Holland's Interest Inventories, and we are finding that there is a considerable intra-individual variability among very bright boys, even among those we are good quantitatively.

Mann: I think John's point is good—how parsimonious are you going to be in interpreting ability.

Bracht: I've seen only a couple of studies relative to the Guilford model. I don't know to what extent this is similar to the ones in special education, but in these studies there is no test of Guilford's model as a total model, or an attempt to develop a program based on the entire structure of the intellect. It was just taking some idea from the models, a very small part, to the extent that Guilford's ideas and the structure-of-intellect help generate some good educational programs. This is fine, but I don't see, or at least I haven't seen, any studies that test or use the whole model for developing things.

Hopkins: It seems to me there are two educational issues here. One, I think, is a good point about trying to establish training programs to improve performance, while the other brings us back to the aptitude-treatment interaction business, and Glenn did a study that I think is quite closely related to that since he defined students according to relatively narrowly defined and purely measured abilities. Do you want to say a word or two about that, Glenn? The PMA spatial relations would essentially be one of Guilford's tests.

Bracht: We had three narrowly defined abilities: spatial relations, verbal ability, and numerical ability. We tried to develop programs which would...well, it was a kind of compensatory model, where we tried to develop a program for students who might be high in numerical ability, but lower in, say verbal or spatial ability, who could then perform well because of their high numerical ability, and in turn, programs for students of high verbal ability. Quite successful in that attempt, but I don't know how to relate it to your discussion.

Hopkins: Glenn completed an ATI study in which he made a considerable effort to match instruction to the different skills of different groups. He had a true experiment. One group had verbally oriented instructions for such skills, while something else was done for the other groups. I think part of the reason that Guilford's model is misinterpreted is that people look at it, you know, in periodic table kind of elements thing, not realizing that covariance is delicate and so is factor analysis. I think most of us have some real reservations about Guilford's methodology. He seemed to rotate to where you were confirming his expectations...things like that.

Stanley: Casa did us a favor, didn't he. Getting us out of the old graphic rotational methods that could confirm your own hypotheses so easily.

Hopkins: Guilford has nevery really left his orthorgonal graphic approach and that is the ultimate in subjectivity.

Mann: Anyone care to make other comments about the use of the factor analysis as a general guide for didactics? If factor analysis in a position to at least provide a legitimate assumptive base from which to proceed, in terms of remediation or of individualization of instruction, I am thinking of Fleishman's work where he believes you should train differently at different age levels for a specific task, because the factorial composition of learning, or whatever, shifts at different ages.

**Feldhusen:** I just don't know an answer. When you begin to relate factor analysis to education, I think of factor analysis only in terms of trying to identify abilities that otherwise are presented by a number of measures, or what have you. When you have taken this big step of relating it to remedial functions that is something else.

**Stanley:** I have been concerned—factor analysis is a nice group technique for certain data reduction purposes, obviously—but it is easy to verify the factors, try to make "thing"-constructs out of them...things out of them that they really don't represent. But they are sort of phenotypical, rather than genotypical, not some underlying thing. Well, you can see that in Guilford's structure-of-intellect model directly in the three dimensions. Another thing that bothers me is, factor analysis is primarily a group technique and individuals vary. I remember using PMA back when Thelma Thurstone came out with a practice book—it's a red book, the green book, and so forth. Do you know about those, Glenn?...where you could train people? Okay. They were cute, though, and I had a young daughter, a preschool daughter, at the time who loved to take tests. (You would expect my daughter would, or she'd have to leave home.) But, she got a lot of fun out of taking these tests so we got her trained on taking PMA—they were meant for classroom instruction, about 20 minutes per page, but once you get a fairly bright youngster to train they can rush through them in a minute or two—she went through all three booklets. And it was clear that at age five, her best ability was spatial, and verbal much less developed than spatial. I gave her the PMA test after this, too. Now later, as a high school-college student, it is clear that her worst ability was spatial and her best ability was verbal. Her asymptoting was going on at very different levels. She went up faster on spatial and asymptoted early; she went up more slowly on verbal, and asymptoted much later—relative to age mates, even. So, I think that this distinctive individual curve for each type of ability during the person's developing life is important. It is like height. Some people, such as I, who were essentially full-grown at a full height at age 12; others reach full height at age 16. I'm pretty sure that in mental things that some people reach their Raven's Progressive Matrices—full height—or most of them, at age 12, others at age 16. Verbally, some people have one type of accelerated curve, and others have another type of curve. So, I am rather always a little uncomfortable with

with general conclusions of groups, when clearly for individuals, we may have quite a different pattern, and for some of the very bright youngsters I work with, we get a reversed pattern. If you give a very difficult college-level test to an eighth grader, you'll find that they will start off getting items wrong, and then end up getting them all right. The reason is fairly simple. I think people say they are rather careless, but that is not true at all—they are not careless at all. What they are doing is, the early items on the test involved specific information based on experience that older people tend to know and therefore get them right. The younger people tend not to know, or they involve very simple reasoning that is too simple for these bright kids—they just can't believe that the item is that simple. As the items get harder, they get harder because they change in factor content. They are depending more on abstract reasoning of the levels three and four in Bloom's Taxonomy, for instance. Now...application of analysis. Now, these youngsters—the bright ones—are particularly good at application of analysis. So, you get the paradox of the youngster who takes the Binet Mechanical Comprehension Test, Form C—the hard test—and misses five items, but he doesn't miss any of the last 21 items in the test. This 13-year-old eighth grader gets them all right. He misses five easier items down the road. Now he is different on that test from another youngster who is older who makes the same score—55—because the other youngster is going to get them all right up to about—out of 60, up to about 50—and then he misses five, from number 50 to number 60. So just lumping people together and doing the factor analysis, I think, oftentimes will cover up a great deal. It doesn't tell you very much about individual differences.

Feldhusen: Your original statement of the question implied that you are looking for remediation, somehow or other, or potential.

Mann: Individualization of instruction—let's put it that way.

Feldhusen: Alright...fine. Now, I've just recently had factor analysis of batteries of tests that included a problem-solving battery which was my major interest. And, let us say, that in the long run I am concerned with improving problem-solving ability, but I am a bit hard-pressed to know how to train that in a global sense. The battery involved in this factor analysis included a concept formation, several concept tests formation, several logical thinking tests, some perceptual tests. Now, it turns out that a factor, and a rather large one, emerges and is represented by my problem-solving test, and by some components of the concept formation tests, and by an element of the logical thinking test, and a couple of elements of the perception tests. Now, this will not be an end in itself, so please don't criticize me before I've had my full say here. But, I propose as a first guess at a development of training approaches, that this factor analysis does give me some leads about how I might go about devising training—even individualized training, here. The ultimate might be to get individual scores on children to detect weaknesses in individuals on concept formation and logical thinking, you see. Do you agree with this as a possibility?

Stanley: At least it tells you what to look at, to some extent.

Hagin: Is your factor analysis just, then, a sophisticated tool to analyze the thing that you are interested in?

Feldhusen: Well, you see that it could continue...well really, I could be done with my factor analysis from this point onwards. I have tests of each of those elements, and if my experimental investigation confirms that by teaching some concept formation and some logical thinking and some perceptual abilities that I indeed do improve performance, and I indeed do want to improve global performance on my problem-solving battery. If I can demonstrate that experimentally, I have no need to come back to my factor analysis anymore. Although there are some reasons why I might want to. But, I don't have an immediate need to come back to it. It has no longer a necessary function in the experimental investigation, that is, the next stage, and certainly no necessary function in eventual application if the system happens to work.

Stanley: I had some brilliant thoughts here, but my attention has been sort of divided...I recall what I was thinking about...I think we have to keep two things in mind, fully. Factor analysis is correlational; it is descriptive; it is not inferential. It does not tell us, for instance, what causes what, since this is a correlation among things, that to some extent like all correlations, limits our ability to confidently predict causal relationships. It may be consistent with the causal pattern, but it is not demonstrating it. Another thing is that in the process of purifying the factors, ala Thurstone, in the 1930's, you run into the possibility of purifying out the element of the criterion that is important. Now, Dwainshaw, in 1949, put to a test Thurstone's and some other persons' fond hope that they were going to improve predictive validity of test for predicting school grade by having pure factors that correlated poorly with each other, but they correlated most poorly with the criterion, too. Dwainshaw tried predicting, I believe it was high school grades, with PMA and he found that you could do no better with PMA than he did with Otis, because the Otis, while it was heterogeneous, had the right heterogeneity for the criterion, whereas only two of the five PMA tests carry any beta weights significantly greater than zero for predicting the criterion. So, in effect we are wasting a lot of testing time fooling around with those three tests that didn't predict the criterion. You are better off to throw them out, and use the two that did, and make them more reliable and sampling block. So, one can—again, I keep citing Keller's 1927 article, Mental Factors of No Particular Importance—get some nice, pure things that don't relate to the criterion. So, in effect, one has to have some kind of factor analysis of criterial measures, too, to look at. Task analysis will provide some idea as to what it is you are aiming for with your tests.

Mann: Ken?

Hopkins: Just a follow-up word on this correlation-causation situation. We are constantly reminded that correlation doesn't necessarily mean causation. Less often, we are sensitized to the fact that most correlation doesn't tell us about causation. Hours of study will correlate about zero with grades in college. That doesn't mean how much you study doesn't influence the grades you make.

## REFERENCE

Horn, J. L., & Knapp, J. R. On the subjective character of the empirical base of Guilford's Structure-of-Intellect model. Psychological Bulletin, 1973, 80, 33-43.

# CHAPTER 7

# DIAGNOSIS, LABELING, AND EXPECTANCY EFFECTS

Overview: Special educators by necessity resort to various means of formal classification. Many, however, have criticized such categorizations as setting up inappropriate biases and expectancies on the part of teachers, child peers, the general public, and so on. Others claim labels are necessary to program in a feasible manner for routine, day-to-day matters. In relation to expectancies, there is, of course, the ubiquitous Hawthorne Effect, which is an expectancy aroused by awareness of participation in an experiment. There are also undesirable expectancies on the part of educators caused by the evaluation and diagnosis. Dissatisfaction with labels is in part responsible for the resource room movement.

**Hopkins:** I would like to make an observation about research and special education. I think the major difference between it and most other ed other educational research is that you have a dozen populations whereas we have but one, and each member of your populations differ in every way that our populations differ among its members, in addition to the fact that they are blind, or they are deaf, or physically handicapped, or whatever. So, it seems that the magnitude and the quality of research that will need to be done with the handicapped might be twelvefold, in relation to what it would be with normals. Each population is so different that one would hesitate to generalize results to one of the other subpopulations.

**Hagin:** When we say "handicapped" we are assuming that this child is somehow physiologically different than the members of this normal population, but as yet, we don't know or we are revising notions about some of these things. For example, there are kids with chromosomal defects, and we are beginning to pick them up in school populations. I was taught many years ago that Turner's syndrome had certain characteristics and we could expect a child suffering from it not to develop very far cognitively. Now, we find this is not always true with individuals. So, I come back to the point that we just don't know the life history—the natural history—of the many things that we are confronted with in handicapped children, in terms of physiological and biological characteristics in special education populations. I do think it maybe justifies the fact that several of the people well-known in special education came, not from education, but from medicine.

**Mann:** Well, medicine has always had a kind of heavy influence on special education.

**Hagin:** So, maybe the aptitudes that we are struggling with and mentioning need to be defined more broadly. Not just in educational terms, but in some of the things that one learns in medical school.

**French:** I think that sometimes we criticize special education classifications too quickly. I don't like the classifications under which special education is funded, but since we are concerned with identifying learner characteristics, we surely are going to have to put a label on them once we get a syndrome. We might have a lot of different syndromes—some of us believe that we need more that is finer classification matter than fewer—so that we know the characteristics of kinds and to build programs for them. Some of my colleagues climb on the anti-classification kick and are speaking out against it, but they are certainly not against careful assessment of where the kid is right now, and let's see where we can take him from that point. In a sense, that is classification.

Bracht: The Hawthorne Effect is generally a positive effect. However, it can be a negative effect, and I think we need to be aware of that, also, in interpreting research results. Oren speaks of the demand characteristics of the experimental subject. But if he knows he is in an experiment, he will try to perceive how he ought to behave, and ask, how can I bring about the experimenter's intended effect? Or, what is the experimenter expecting of me. The subjects in experiments may have different perceptions of what they are expected to do, and in that way their behavior may be influenced by their perceptions, rather than what the experimenter intends. Or, sometimes, there may be a general negative reaction if the subjects in the experiments see themselves as guinea pigs, in some sense, and resist cooperating in the treatment, and that may bring about a negative effect.

French: As I try to recapture what Jay was saying earlier, I thought he said there were a variety of characteristics of the Hawthorne Effect, and the only two that I can recall: one is the expectation that people will do well because of their having a new treatment, the other—

Millman: One was the novelty of it...

French: ...the personal interaction?

Millman: The other was a feeling of just being given special attention. Another, was being made aware that you were in the study, which I think was more in line with the original Hawthorne situation. There was a fourth one, but I forget it for the moment.

Feldhusen: Well, there is a very closely related problem and that is the expectancies generated, or the effects generated, by the preliminary assessment. Friends of mine in a guidance clinic have pointed out that in lengthy diagnostic things that might go on for a couple of weeks, that unusual amounts of remission frequently occur. So, it might be a larger problem. We, of course, worry about Hawthorne happening just with the treatment effect.

Mann: When I was working in a child guidance center, John, the diagnostic effect was a common one.

Feldhusen: There is also the possibility of a negative diagnostic effect. I was alarmed over some of the data that we generated in a study back in Wisconsin with Mr. Klausmeier quite a number of years ago. A lengthy diagnosis by clinical psychiatrists following some time later by another lengthy psychological diagnosis-produced data, in which the number of cases of seriously disturbed children went up appreciably. Instead of the possibility of the first lengthy diagnosis having a positive effect, we suspected that somehow or other, it may have had a negative effect on some appreciable number of subjects.

Hagin: There was a negative effect generated by an attempt of a Scottish study, which I can't identify clearly, to identify obstetrical risk early. These mothers went to the National Health Service in Scotland and had been warned that their children were vulnerable, and the amounts of prediction were realized to a greater extent than justified. The mothers' negative concern appeared to be responsible.

Hopkins: I'd like to point out that just because there is a possibility of the Hawthorne Effect, doesn't mean it's reality. I see many people act as if the Hawthorne Effect always is operating, and asked whether we've got data that there is no Hawthorne Effect in a study. I'm not trying to say that the Hawthorne Effect doesn't exist, it obviously does in some situations. But it seems like there's been an oversell in education about the Hawthorne Effect. Where our treatments are of longer duration, particularly with cognitive variables, I think those kind of situations are must less likely. I'm just speaking logically, not empirically, to minimize the Hawthorne Effect and, of course, one wants to take whatever steps he can to minimize it, if it does exist, than have this unobtrusive measurement business. When it's possible, there's no point in telling the experimental group you're the experimental group, and you're to be compared to the control group. That sort of thing is going to inject a potential reactivity in the system. Personally, however, I think it's very unlikely that with cognitive outcome bariables, where the treatment is of long duration, that the Hawthorne is going to contribute very much. That's a personal opinion.

Feldhusen: I want to add something about labeling since this arises so much in research in special education. Now, I'm constantly being bombarded with the notion that if we dare to make predictions and label anybody that the self-fulfilling prophesy will inevitably become operative. I agree this is possible, especially in special education research—that of where the label may lead to the identification of youngsters for a particular diagnostic group. But I'm impressed with the vehemence of the assertations about labeling effects that are coming at us constantly.

Proger: Parents have been attacking the schools about their labeling practices; they have become unpopular. In fact, the resource room movement represents, in large part, an attempt to reduce labeling.

Hagin: Johnson assembled 20 or 30 fairly good studies which could not justify the expense and inconvenience of special classes for retarded children. Then he went back to understand why they didn't make any difference; the only thing that he felt help up was the fact that the teacher's expectancies were limited by the fact that the children had been diagnosed. So, it's not just the parents' complaints.

French: I don't know how carefully Johnson analyzed his data. There have been other studies suggested by Johnson. However, these are mitigating circumstances. I was called in to help with the educational program at one of the state institutions and one of the arguments was that the IQ level of the kids was within the educable range and this would be a good training opportunity for our students. But, in fact, we found that even though they might be of the same intellectual level as children in day programs, they're just a completely different group of children with different educational needs, different behavior needs, and so on. Those that are referred for special class placement are very likely a different type of mentally retarded children than remain in the regular classroom.

Mann: This was Kirk's point by the way, in that book by Helen Stevens. Comparisons of kids in institutions to those remaining at home. But, really, the one referred to in institutions is the acting-out, hard-to-manage.

French: I think the same thing happens with many of the special classes. We might be examining kids that are referred by teachers because they're problems, and other kids who are equally low intellectually aren't referred by teachers because they're nice, well-adjusted kids who get along in the classroom and make reasonable progress, according to the teacher. So, they're never referred. Until we get into some studies where we have a very effective screening program and a random assignment to groups, I think we're without a lot of evidence on the efficacy of special class placement.

Hopkins: I'd be very surprised if one of those studies was not a "magic-type" study, in which case there would be a regression phenomena creeping into the picture.

Feldhusen: Coming back to labeling just for a second. I guess that seeking further confirming evidence of the effect, it would be desirable to try to find ways to avoid labeling, even though Pygmalion in the Classroom is seemingly discredited, in spite of the discredit of the Rosenthal research. In special education research, substantial efforts to avoid that would be desirable—to avoid labeling.

Mann: It is a contaminating variable, or set of variables, for sure.

French: We can, of course, work with teachers to disprove the fact that their kids are mentally retarded. Years ago the situation was brought to our attention that on repeated measures of kids in special class placement, the IQs went down and the teachers were saying, I really try to make things go right and to help these children, despite that the IQ is lower. At the time, they were still using ratio IQ with Binet and most of these kids were measured at when they were six, when the standard deviation was the smallest, and again, at nine, where it's at the median, and at twelve, where it's the largest. We just had conversion tables to use deviation IQs, where there was no progressrive decrease in mean IQ, with the ratio IQs there were. I would think there would be some strategy by which one could direct the teacher to disprove the classification. Surely, there are enough teachers around who would like to believe that.

Millman: Expectancy as a construct doesn't really exist in something that you can touch. So, if there is an expectancy effect, it's really based on the expectancy that the teacher then behaves in a manner different than she would have behaved, had she not had that expectancy. For example, putting the child in a special group, or if she's told that he's not going to make very good progress she may put him in a slower group than otherwise, and therefore, the child doesn't get exposed to certain kinds of instruction. I'm not agreeing with John, that I think it would behoove researchers to minimize the labeling things and to look very hard at how the child is actually functioning.

Feldhusen: Despite the unhappiness of Pgymalion, there's the work of Howard Spiker, essentially in the special education domain, is very positive for short-range assessment of expectancy. Using the very bland situation of telling the teacher simply that here is a child below average in IQ, then giving her a folder in which she sees, indeed, that the IQ is 98—if she looks. Subsequently, watching that teacher's behavior or that teacher working with a given child, Mr. Spiker discovered that her rate of presentation of new material to that child, whom she first was told was a below-average in mental ability, there was an appreciable lower level of presentation of material to such a child, than to another child about whom she's told he is above-average in IQ, and the folder which she gets shows an IQ of 102. Now, this is very short-range research. It's looking in almost immediately after the report, but the effect is a very positive one.

Mann: Elsa Flemming and her colleagues didn't find any long range effects along these lines. The teacher may be changed for a few days, but after awhile she responds to the kid as he is. I want to raise the question, at this point, as to whether expectancy effects are properly Hawthorne Effects; if, indeed, they exist. They, of course, affect the experimenter who then affects the kids.

Millman: Sounds like a definitional question. We usually think of them as different in the situation where the subject under study is a student and the expectancies are on the part of the teacher. The nature of the bias is the same kind of bias, in that you will attribute your results to the wrong thing.

Mann: Let me ask another question—naively, I'm sure. Would it be possible to establish that Hawthorne markers or intervention markers would somehow give us quantitatively measurable "Hawthorne Effects," if you will? Let's say, again, we're doing physical education training where you test for improvement in those areas, and you throw in arithmetic testing, as well, without special training instruction, ignoring the practice effects around the arithmetic test—you get a quantifiable "Hawthorne" increment which you can subtract to give a more clearly defined treatment effect.

**Millman:** The judicious use of control groups of different types could help you separate out these various influences.

**Mann:** But not in terms of the effects of actual treatment measurements on this sample. I'll repeat, you're teaching physical education, and you're giving him arithmetic testing along with physical education testing. As the arithmetic rises, you assume that it is a general Hawthorne Effect which you then can subtract from the physical education effect.

**Millman:** I feel that's a possibility.

**Feldhusen:** One could, of course, in a longitudinal study, get a crude way of assessing novelty as Hawthorne Effect. Novelty ought to wear off sometime or other, and if one can compare initial performance—when a new treatment is first introduced—with the way performance, let's say, degenerates, after prolonged use of the treatment, why, I think, you would have a way. Now, the thing I'm talking about has been used another way. People have suggested that it is not appropriate to bring in an educational evaluation until you've gotten into a stage that the treatment is well established and no longer new and novel. So, this seems potentially a way of getting a hook, if you're asking, how can I get a measurement of how much a Hawthorne Effect is. If I could compare, somehow or other, preformance right at that early stage when the new comes on, with the way people are performing after this stuff has been around for a couple of years, why that ought to be the differential that tells me, in part, the answer.

**Millman:** Well, on a scale from 0 to 10, I'd say—I don't know how this scale business get started. I picked this up from my kids. I don't know where they picked it up. They say, well, Mommy, and Daddy are going out; we have to get a babysitter; how about Susie or somebody. Jeff said, on a scale from 0 to 10, with 0 meaning not like, and 10 meaning like very mcuh, she's negative 3.

**Hopkins:** Well, another aside—we've got a weatherman who gives the weather that way. He reports what the temperature is going to be and all that. Then he gives you the general IQ parallel; that is, on a scale from 1 to 10, tomorrow will be a 6. It's kind of interesting. He doesn't define the influence.

Millman: What do you think the underlying dimension is?

Hopkins: Good and bad weather—what else?

Feldhusen: I guess I'd like to go back and ask why I don't get 10 for my suggestion. Supposing I have a treatment going on in a school system, and I get this first year a mean performance 79; and next year I discover the mean performance is 69; the next year, then, it continues 69 thereafter.

Millman: I think it's worth looking at, and so on, but I think it's not perfect data. You've got differences in scale units and artifacts of that type creeping in. You know, a whole mess of things that make it so that it isn't clear-cut and perfect evidence. But on reconsideration, would you settle for 8.3?

Mann: When John and I publish our equation, you'll be sorry.

Hopkins: I'll give it a very high rating.

Feldhusen: I'm sorry I brought this up.

Hopkins: Just another comment about labeling. Of course, words are a means of communication and humans by very nature are parsimonious and all communication is imperfect. When we want to convey something, we want to convey it as precisely as possible. That's one thing; but then, certain groups are in continuum. There becomes a stigma associated with a term and as one term begins to develop a stigma, a new one is created; we all know that. In the old California legislation, they didn't send children to MR classes; there were .1 classes and .2 classes, and nothing sounded more innocuous than that. But in time, they did become stigmatized, and the resource class, too, is going to become a stigma.

Mann: You are, of course, right—whatever name you give eventually sounds plagiaritive if the associated children have negative qualities.

French: Well, the name really changes people's percetions—certainly only to a very slight degree—and expectations, too, if a teacher sees a kid always at the bottom of the class— whatever you call him. What is the expectation of that teacher? The kid really identifies himself, in many instances, and therefore, the identification is based upon other factors than the label given to him. There is also a self-labeling effect.

Hagin: Well, I'm not sure...We have found that teachers are pretty poor at picking out the kids that need mental health services. For instance, we asked—and we do it every year—which children might they predict would have learning problems. I remember clearly the first year every name on the list was either Spanish or Chinese, and middle-class children were missed completely.

Mann: Was this an adequate reflection of the instructor?

Hagin: No, because we had many middle-class kids who had problems, but they were not perceived this way by the teachers. We also found that boys, if we use clinic populations for studies, boys came out many times. When we study the intact groups we found that a surprisingly number of girls who probably wouldn't be referred to the clinic, were identified as having learning problems. So, certainly in our samples, we whouls be aware of the bias, in effect, of teachers' judgments.

Mann: Any further additions to make to the topic?

Feldhusen: There may be a positive side to this, too. I have been amazed, over the years, at how anxious and enthusiastic subjects from the Termins group of gifted have been to point out their mentality. There were several people at Purdue who announced—very frequently—that they were among Termins gifted subjects.

Mann: Well, I remember one of the two who had been one of Termins gifted children, and who apparently could not tolerate being the super genius that he had anticipated he would be. He was bright and successful enough as it was. His case was eventually a tragic one. His expectancy level may have been too high; he just couldn't match up to that gifted evaluation.

Hagin: There are also parents' expectations. There are certain diagnoses that parents like, and will go through some trouble to get their child diagnosed that way. Not only because of the facilities that are available, but because of their own needs. It is better to have an emotionally disturbed child—a schizophrenic—in New York, than to have a mongoloid. The mongoloid parent is in the lowest of the packing order. For a while it was important to have a brain damaged child, becuase it becomes not one's fault so much as maybe some of the other things. People will shop around in order to get the right diagnosis.

Mann: Learning disabilities is by far the preferred one right now. Dyslexia is beloved, too.

Hagin: This is the disease nobody had seven years ago, and now everybody's got it.

French: The court ruling in Pennsylvania states that all mentally retarded children should be considered brain injured. I've lost the phraseology of it, but it was primarily to make them eligible for all the benefits that occur through special education for brain damaged.

Feldhusen: Again, to relate this to a research more directly, I think that if diagnostic classification and the labeling existed before one ever drew a sample for treatment, I'd be less concerned than I would be at the situation where the research and subject selection was based on diagnosis done within the study and the subject then is classified and labeled, and the classification, or labeling, fixes him to a treatment condition. It would seem to me that this is something to be worried about, perhaps more so, than in the situation where the diagnosis in the classification has occurred for other reasons apart from whatever research one may be doing.

Millman: I don't follow that. I mean, I hear what you said, but I don't see the difference.

Feldhusen: I'm saying that if a treatment group is created as a result of your identifying—your knowing him as a mongoloid, and having him assigned to a treatment group—then this seems to me the most potentially troublesome situation as relates to a research setting here, in terms of expectancy, positive or negative. But, if the classification existed before and was not a part of our research perse, then at least a new label and a new diagnostic classification is not suddenly descended on the child and his family. That might have an overwhelming effect on the study—the impact of, for the first time, categorizing and labeling.

Mann: Most children are a priori classified, I would say, prior to study. Now you can, of course...in some of our research programs we are attempting to establish new groupings by characteristics, but by and large, the children are coming from identified populations, that is, one already identified as handicapped. For example, Sabatino, one of Joe's faculty, is trying to identify the children who are handicapped by their learning characteristics, if you will, and to use these, I think,as predicters in the scheme of things, while ignoring MR or BI classifications, and so on. But I don't know...this is not the case, by and large, in research.

Feldhusen: Well, I would think that the re-labeling, as it was, would affect the subjects within a treatment-type study. Different types of expectancies would be generated, and this would possibly affect the results of the study.

# CHAPTER 7

## REFERENCES

Brophy, J.E., & Good, T.L. Teachers' communication of differential expectations for children's classroom performance: Some behavioral data. *Journal of Educational Psychology*, 1970, 61, 365-374.

Dangel, H.L. Biasing effect of pretest referral information on WISC scores of mentally retarded children. *American Journal of Mental Deficiency*, 1972, 77, 354-359.

Finn, J.D. Expectations and the educational environment. *Review of Educational Research*, 1972, 42, 387-410.

Hammer, B. Grade expectations, differential teacher comments, and student performance. *Journal of Educational Psychology*, 1972, 63, 454-458.

Harris, C.M., Martin, R.R., & Haroldson, S.K. Punishment of expectancy responses by stutterers. *Journal of Speech and Hearing Research*, 1971, 14, 710-717.

Holmes, D.S. Effects of grades and disconfirmed grade expectancies on students' evaluations of their instructor. *Journal of Educational Psychology*, 1972, 63, 130-133.

Kester, S.W., & Letchworth, G.A. Communication of teacher expectations and their efforts on achievement and attitudes of secondary school students. *Journal of Educational Research*, 1972, 66, 51-55.

Marchall, J.C., Watson, E.P., & Sokol, A.P. Dimensions of teacher expectations and student behavior in high school social studies classes. *Journal of Educational Research*, 1972, 66, 61-67.

Powell, G.E., & Watts, F.N. Determinants of expectation in imaginal desensitization. *Perceptual and Motor Skills*, 1973, 37, 246.

Rothbart, M., Dalfen, S., & Barrett, R. Effects of teacher's expectancy on student-teacher interaction. *Journal of Educational Psychology*, 1971, 62, 49-54.

Salvia, J., Clark, G.M., & Ysseldyke, J.E. Teacher rentention of stereotypes of exceptionality. *Exceptional Children*, 1973, 39, 651-652.

Schwarz, R.H., & Cook, J.J. Teacher expectancy as it relates to the academic achievement of EMR students. *Journal of Educational Research*, 1972, 65, 393-396.

## References

Shriberg, L.D. The effect of examiner social behavior on children's articulation test performance. Journal of Speech and Hearing Research, 1971, 14, 659-672.

Soule, D. Teacher bias effects with severely retarded children. American Journal of Mental Deficiency, 1972, 77, 208-211.

# CHAPTER 8

# EVALUATION OF CURRICULUM AND INSTRUCTION

Overview: Accountability has achieved such prominence in the applied realm of educational research that it has begun to assume the status of a discipline of inquiry. Indeed, a large department was established at Ohio State University for just such a purpose. Yet, in the day-to-day real school world of the special educator, accountability has remained a nebulous and often threatening concept. This chapter attempts to put the topic into perspective.

A major source of confusion over accountability lies in what is being evaluated. To some, accountability means the evaluation of the performance of an individual teacher, administrator, or other staff member; while this aspect of accountability is touched upon in this chapter, the main discussion occurs in Chapter 9. To others, accountability means evaluation of a given school program's effectiveness averaged over all teachers and all pupils; this is primarily the focus of the present chapter. In other words, one is interested in total program evaluation.

In program evaluation, one must distinguish between evaluation of only instruction itself, or the entire curricular Gestalt of instruction; the latter is far more inclusive than the former and might conceivably tap such far-fetched areas as extra-curricular activities, physical facilities, and so on. Apart from which definition of total program evaluation he accepts, he must make decisions on what evaluation schemes are to be employed in data collection. One might utilize standardized pre-and post-testing within a given academic year and hope that he has some appropriate criterion or expectancy level of gain. Or, one might use only one testing session in the current year and pit its results against those of the preceding year (when a different program was used). Or, one might want to use the new-breed testing of criterion referenced measurement (see Chapter 11). Or, one might wish to engage in cost-benefit analysis. Or, one might wish to bring in external auditors and/or on-site evaluation teams. Or, one might want to use descriptive observation as a type of systems analysis of his project (see Chapter 14). Finally, one might wish to employ any one or more of the above options under a special arrangement known as performance contracting whereby a commercial company manages the school's instruction on the basis of a money-back guarantee/piece-work basis. Just as with formal experimental design, the possibilities are endless, and this only adds more confusion to the muddled accountability movement.

Feldhusen: Researchers in conceptualizing have problems that generally do not concern themselves with what kind of practical realistic long-range role you might have for a reading deficiency, or a reading problem, or what have you. People do address themselves to this task very often but yet researchers may neglect it; I myself don't know for sure what the problem might be for a child who is deficient with a certain amount in reading. I don't know possibly anymore about that than the fact that it shows up as a score on a test or as a discrepancy from a norm. Stufflebeam's CIPP model might be useful for a researcher to apply here to get greater understanding. In total context, I am saying the researcher really doesn't know what deficiency means and that I, for one, am ignorant of what the deficiency of, say two years in reading, means contectually in the life of a child and then, number 2 - I get to the second stage of the CIPP model; what kind of a good goal ought I to have, other than simply trying to bring his performance up to improving, somehow? What ought I to be trying to do for him that would be worthwhile to do for him? I think that such considerations might often, for the researcher, be important issues to go back to. I fear that as researchers, we often get trapped into extremely narrow conceptualizations of what we are up to. Here, the practitioners who first uncover and clarify the problems, ought to perhaps make better input to us to help us understand more. I think this is a criticism that probably is inherent in Mr. Stufflebeam's work. Let me add that it is very well set forth in a paper by Henry Gideon -- you know from the Office of Education - that educational researchers have perpetually been trapped - he asserts, as does Stufflebeam - into such narrow views of what they are working at that the large picture of what they should be trying to do to help kids, or improve education, gets lost and even in much of our discussion, I am lost that way. As we talk about it, shall it be perceptual training or shall it be specific training in handwriting, or what have you? I still am not sure what you might have in mind, as to what you want to have happen for the child and what he is supposed to come out with.

French: Are you saying that such objectives as improvement in word recognition skill is too narrow for a researcher to focus on?

Feldhusen: No; I assume that improving the academic and functioning is a major focus for us. I am asking: what do I want to do for a poor reader and as this is a problem as a researcher that I am concerned about. Why do I really want to be concerned about improving his reading?

| | |
|---|---|
| French: | Is this a research issue or just a philosophical goal? |
| Feldhusen: | I think it would simply help me, or any other researcher, understand better what he is up to and what he is truing to aid or abet if he already understands the goals of the practitioner. I am stating faith in Mr. Stufflebeam's CI parts of the CIPP model here. |
| Proger: | Well, why should this be important in a research... If a school administrator said, "I want to have my youngsters in an MR class reach 5th grade competence" that is the task - Who? |
| Feldhusen: | Because you may trap me into helping you design and carry out a problem that is ill-conceived. |
| Mann: | That is a possibility, but I don't think that the researcher necessarily has to pass judgments when providing assistance. |
| Proger: | Are you saying that the researcher has a judgmental responsibility to discuss values before proceeding to design a study for someone else and collect data? |
| Feldhusen: | I have the feeling that if I were helping someone design, that I could help him design better if I understood more adequately the context of the problem and a proper way of conceiving the goal toward which he is working. Now this point is most forcefully set forth by Mr. Gideon in his paper: that educational researchers have persistently allowed themselves to operate in such narrow framework that they never really understand the context of school situation or however you would pronounce it; and secondly, that very few people sit around and think seriously about this and I guess Mr. Silverman echoes this point in "Crises in the Classroom", too -- we don't know how such larger concerns relates back to designing research, but I am concerned that possibly I persist in designing poor, narrow or inadequately conceptualized research because I have not addressed myself adequately to the larger issues of what really is the problem and what the sensible goals are for educational research. |

Millman: Well, anyone who has been an editor or is advising students on these problems can certainly give many anecdotes of situations in which the problem he was really concerned with was not what the student thought it was originally. The need goes back to thinking about the problem. As I think I said earlier, I am, for that reason, talking about what are the questions in special education.

Mann: There are a number of such questions, of course. For an example, "How do you restore this kid to normal status in some fashion?" It is one that may not be stated openly but it is implicit in many of the procedures we employ to special education. Of course, the spoken or unspoken question of parents - and implicit in many of our technologies in special education is, "Are we going to label him normal through education?"

Millman: I'd say the philosophy issued, John, really should perhaps be associated more with the feelings of say Michael Scriven who talks about evaluation. I don't see as big a role for values in research. Certainly Michael Scriven would say, about evaluation, that you have to go back and ask questions such as you are suggesting.

French: He would say there should be an independent assessment of goals. The project director might set certain objectives to accomplish and, if we limited ourselves only to an evaluation of his objectives, we may be doing a disservice to the total educational picture or the total life space of the individual. Scriven would say the evaluator must also look at the expense incurred in obtaining those goals: that if one sets only reading improvement as particular global goal and sees that reading has been improved, at what expense has he achieved that goal; have the children, as a consequence, platuaued in arithmetic or social adjustment in the classroom; at what expense has such a program been successful? Evaluators do seldom look at that. They only look at the specific objectives that the principal investigator enumerates at the beginning of the study, rather than looking at things from the more philosophical point of view, what all should we evaluate as a particular sub-portion?

Mann: For an example, we used to give you a choice, perhaps in teaching an EMR to read at a level approaching his limits, of having a nervous EMR child who can read or is a relaxed one who can't. This is a critical decision, but is it a research decision?

French: I don't think the researcher should make that decision but a comprehensive evaluation might give you an answer.

Hopkins: A bit of irony here. The latest paper I've read by Scriven called Goal Free Evaluation... and they already know more about goals than they need to. Let's forget about setting goal objectives and be more explicit about what we do.

Mann: Just go ahead and take the youngsters and see what works?

Hopkins: See what the outcomes are... whether they're pros or not.

Feldhusen: Or see what any outcomes are there. Anything in sight.

Bracht: I didn't quite interpret his book the same way you expressed it, Ken. I think what Scriven is advocating is that the evaluator should not be bound by the goals stated by the project director. Rather the evaluator should decide what data to collect for an evaluation. In that sense, he is talking about goal-free evaluation. I think he has a poor choice of terminology here. I don't think Scriven at all intends that we have goal-free evaluation. I think what he is advocating is that the evaluator be free or not bound by the goals stated by the project director. I think at some point, though, there will have to be goals or values by which we judge whether the outcomes are desirable.

Hopkins: A little anecdote he gives, Glenn, you will remember that the evaluator comes into the situation when someone in the project starts telling him about the objectives and he says, "No, I don't want to hear those. I know it is a reading program and that is all I need to know." So, he may be overstating his position in a theatrical sense but that is what is there.

Mann: You are saying that the evaluator should go off on his own tack and evaluate what he wants to evaluate. You are saying that people evaluating the effectiveness of the program should examine other behaviors and goals attained besides the ones specifically focused upon by the project's personnel.

Bracht: This is an external evaluator who is going to collect some data-- Scriven doens't want to collect the data that the project director tells him to collect. He wants to make his own decisions about what to collect.

Proger: Who then puts the ultimate values on the importance of the findings? Is that up to the evaluator, or is it up to some additional external agent? Who has the responsibility for interpreting the importance of the findings - statistical and pragmatic significance, etc.?

French: I thought that was the point that Scriven was trying to make - that the data was contaminated if the principal investigator played too strong a role.

Proger: In interpretation?

French: Well, yes, because there he really didn't collect enough data.

Feldhusen: Are you talking about the evaluator now, or the principal investigator of the project?

Frence: If the principal investigator specifies what the project is to be, and how it is to be evaluated, I think it raises the question as to whether or not there has been enough evaluation.

Proger: Who decides what is the value of the findings by anyone?

French: Well, society has to put that value on the findings.

Proger: Who is the most qualified interpreter?

French: You and I.

Feldhusen: Well, Mr. Skager is notorious for arguing, of course, that the evaluator should be utterly free to pass judgment upon the results of his evaluation and, of course, protagonists - one protagonist on the other side is Mr.

Stufflebeam - he argues to the contrary. I don't know where the battle stands at the present time...or how to resolve it. I guess my faith is with the Stufflebeam position; that the evaluator should be an objective information producing agent and that the judgment function as to values occurs by the decision-makers who receive the information.

Mann: Would someone bring it back a little closer to exceptional children or in special education for a final note.

French: I think this applies to special education more than anything else, and I'm - I don't know who I agree with - Stufflebeam or Scriven but I think it is very important that in evaluation we not be contaminated by a very narrow view of any particular investigator and that as we really think about the bigger picture of evaluation, we not be concerned with only the very specific aims of a particular activity or project.

Mann: I'd like to resolve some more issues on teacher acceptability. Should we go, for example, by using measures of standardized achievement tests at the close of the year or by a CRM system as we are using and so on, are we being redundant in looking at accountability from this standpoint?

Millman: There's a paper that your audience ought to know about that's called, "The Teacher Accountability of Three Levels." It addresses itself exactly to that question, although not specifically in special education.

Hopkins: Just a related point. In Colorado, achievement measures are used in assessing accountability of the district level. Districts are given percentile ratings in terms of their student performance with and without adjustment based upon sociological characteristics of the district.

Proger: I'm interested in what Ken brought up here. This is the type of accountability I was thinking of. Just what would school systems do feasibly for the total-year accountability? To say the district has achieved its purposes or hasn't? Just go into a regular standardized test or what?

Millman: Well, I think it depends on accountability of what. If you're talking about accountability of a district or a school system or the accountability of a teacher. I think there's a difference in what you're doing depending on the level that you're talking about.

Hagin: Could I move along a little bit from accountability which implies a sort of either/or to one that may be a point that might have more application, in terms of research, which is cost effectiveness. Where there are alternatives- (and we've discussed some designs that we could offer as alternative treatments) - and where its not a matter of either/or but of selecting a number of treatments. That one which for the amount of money, convenience, and so on, would offer the best results.

Mann: Well, one of our major problems, I think, in special education research and this sort is the fact that we are not alone or only concerned with the curriculum accountability. There is the managability of the child - his emotional status. The parents' satisfaction is very often a deciding factor in determining what programs go and don't go. Far more than in regular education, a parent may say, "I like this teacher. The child is happy with her. My kid can see other kids in the lunch room and go play with them once a day." Such factors may wash out academic and training criteria.

Hagin: But should our decisions in special education be made on the words of the more vocal parents? We are faced with some choices when we have research where optimal things are done, Heber's program for instance. Is this too much investment of time, energy, money, pupil time, and so on?

Mann: Well, I think, maybe from the previous conference, I haven't read the results from a highly founded newspaper. He's taken two groups of inner city young preschool retarded youngsters (black) and they've had massive omnibus intervention programs with one group. Of course, the other group is the control. He reports rather dramatic cognitive growth as per assessment versus intelligence testing -- 30, 60 points, 130 IQ's and so on.

French: I think they selected them on the basis of the mothers being of below 70 IQ, and then they worked with them as babies,and I was wondering what instrument they used to show that IQ growth in the below 3 year level. The study is difficult to get hold of.

Hopkins: I did talk to someone who had seen a film that Heber had apparently produced. What struck this person was how closely related the things that they were doing were to what goes on in an intelligence test. I think the real payoff will be grade one. Whether they read better or do anything of social value better. After all, the score of an intelligence test isn't transferable in some general sense value.

Mann: The cost, by the way, was extremely high.

Hopkins: $50,000 --this is newspaper reporting. I've figured out there's 20 experimental subjects or 25, and it's 500,000 dollars or something like that.

Mann: Now we know. So there is a cost effectiveness factor, obviously.

Millman: It's interesting though. I mean,if Heber can get these kinds of data,you know,and results that will last and transfer to school-related subjects;even at that kind of expense it's worth knowing. Then we can sort of see what can we do without,and where can we cut corners. It may be that we can come up with something economically feasible.

Feldhusen: I think this is often a problem that from the consumer's point of view when looking at research, we want some kind of practical application very quickly. I think this point is being very well made here,and a Heber's study, that while the continuum from basic to final application in some real setting is hard to define, nevertheless, it would be wise to grow appreciative that could be work going on at a level where it is still quite ridiculous to talk about practical applications. Cost effectiveness is just an irrelevant issue here, really, where one is exploring and developing a way to stimulate early cognitive growth.

Yet that does beset us at all times. I cannot do research at Lafayette, Indiana, or at Indianapolis, Indiana, unless in the application form I have to specify what the outcomes will be for the school system. Not even for the world in general but how Indianapolis or Lafayette will get a practical payoff. I think this represents a terrible injustice to the researcher who might be working quite toward the basic end of the spectrum.

Hagin: Well, I think we have to accept the fact that what we do in special education will be more expensive, and we'll probably have different relationships to accountability than general education. It is important certainly to know to what extent we can apply without sacrificing results. This is why I'd like to stay away from strict accountability and to use the term,cost effectiveness. We found,for instance,our nursery program took almost as much time in conference and supervision than was spent with the children. We are willing to justify this because of certain results that we found. But,I think that it's important to have some basis for backing up the high costs.

French: Well, I wonder if you're not overly optimistic about how we can cut costs. Higher Horizons Project in New York City was able to show effectiveness in one school with a lot of money but when they started cutting back, it washed out completely.

Mann: But this is true in many places with many projects that don't reduce costs. So you have to wait and see.

French: I think this has relevance to some of the things we're talking about in accountability and teacher effectiveness and the criterion measure and performance contracting,where we find that the instructional program becomes one of teaching for the test. I can recall years ago,visiting the classroom where the teacher had information items from the Wechsler as the topic for consideration for the class. What are the four seasons of the year? He was drilling them on it day after day,after day. We have to be very careful that the instrument we use to assess growth won't be the very item that the teachers train for. It's certainly has been true in a number of school systems.. Teachers have gotten the standard achievement tests and really drilled the kids on those standard arithmetic problems and those particular reading comprehension passages,and they really look great when the achievement tests come in at the end of the year.

Feldhusen: We have many reservations about that post test assessment in research setting, particularly, and about their relevance to other assessments. The mastery teaching movement, for example, has gone completely gaga with the notion that when a child has demonstrated 90% proficiency that you've made it and once and for all. I think the lesson out of this is we need many more instances of people sticking with it, both from a longitudinal point of view to see what the long range course of an initial effect is and in some ways extend it, as Jay's suggestion in his earlier concept of multiple criteria. That implies transfer-oriented criteria. Criteria that perhaps move toward more significant or practical application, or use of whatever the skill is or it's relationship to subsequent learning or performance. We're still grossly beset by the researcher administering a single, one-shot post test exactly at the end of his treatment; after the last portion of the treatment is done, he administers his post test and goes his way and never makes any further assessments whatsoever. I think this must certainly characterize a large portion yet of what you as, an editor, see of experimental studies. Is this not so?

Mann: Yup! The professions still count the publications, you know. At an APA Convention some years ago, for your job application they'd ask you to list the number of publications, not what they were or how good they were, just list the number. Obviously 20 is better than 5. Anybody want to talk about performance contracting for a moment?

Proger: Just one nonresearch comment about performance contracting for the retarded. One thing that seemed to shake most people up in the school district of Philadelphia was that it never was communicated to the teachers adequately and it implied teacher evaluation when it should have been pointed out that they were not on trial and that we were looking at the program as a package trying to evaluate it. This created a tremendous amount of hostility on the part of the teachers. I'm sure, from his point, most of them sat back and didn't bother cooperating.

Hopkins: The whole strategy using performance contracting is to insure that a lot of apparent gain is going to take effect instead of a regression situation. A real good example is about 4 school districts that went into a performance contract arrangement, along lines of the Texarkana project. This fellow Thonet was able to convince them to do, in effect, an experimental study. So he got a control group and all that, and sure enough, there was no difference between those who had been in the program and those who weren't. Of course, this type of result is not generalizable to other kinds of instructional strategy, but it illustrates how naive many decision-makers are about measurment and validity and things like that that are really important terms in making wise administrative decisions.

**Millman:** From a research point of view, I think the best advice I can give in order to measure the effectiveness of a program that's been contracted for is to have, as the base of comparison, what happened the previous year before the school started the new program. In other words, the assumption is being made that this year's second graders or MR children probably isn't much different than last year's. See what happened to last year's group. How much they gained or did not gain or so on. Whatever measures were in use in that school the previous year and then compare the new groups to what happened to the previous group. Now if not everybody is being trained under the performance contract, then obviously you can have random assignments experimental control group. But even if that isn't possible, I certainly would say use as your measures what you said was used the previous year. In other words, you have some idea of expectancy. If you merely look at pre-post test kind of scores you have all the problems that Ken alluded to.

**Proger:** What happens in the situation with retarded where you're lucky if you even have a post test measure of achievement? Let's say you're dealing with educables. Many school systems don't even bother to even test these kids for regular achievement on a standardized test of any type. What happens then?

**Millman:** Well, what are these contractors supposed to be doing?

**Proger:** They guarantee a year's progress.

**Millman:** In what?

**Proger:** Reading and arithmetic, usually.

**Millman:** You're telling me now that these mentally retarded children are now going to be under a system supervised by the contractor, would have no reading or arithmetic information about them and yet...

**Proger:** If you're going to use your paradigm going back to the previous year, I think ..

**French:** You could start your contract in the spring instead of fall.

**Millman:** Well, that's right. You bring them in in the fall and it seems to me that in the spring you have some kind of information. That the comparison is not how much the kid is getting from spring to fall, but how much he is, compared to where he would be expected to be in the regular system.

**Proger:** How valid, though, is a once-a-year measure as an indication of presumed gain?

**Feldhusen:** I see a little lesson in this that might be imparted to the world at large, that has not yet been beset by the accountability onslaught here, which is to get going with some of your own assessments so that in case you're threatened next year, or the year after, you will also have some of your own data, to wheel out contradiction to what might later be produced by the contractor.

**Mann:** Ken? I want to take... there are just three minutes left here. I want to bring up a very brief question - the question of curriculum calidation.

**Stanley:** Oh, boy.

**French:** I thought that was your topic for the next conference.

**Millman:** Do you really mean curriculum evaluation or do you mean instructional evaluation?

**Mann:** Well, we went over instructional evaluation two days ago. So it is curriculum evaluation.

Millman: You mean where you are looking at the values and objectives and so on? Well, I think that the contribution that has been made in the measurement area in this topic of curriculum evaluation has been to alert the profession to the necessity to consider more than just standardized test results. Bring to bear in the evaluation all the kinds of information that effects the decision making. You know, that is crucial for the decision making.. This involve things like goals in community. It involves cost. It involves affective measures. It also may involve kinds of things that standardized tests might pick up. But clearly, I think, the role of standardized testing - cognitive kinds of things is - has been seriously deemphasized on this writing and I think that anyone who is interested in this should read the references.

Mann: Glenn?

Bracht: There are both logical analyses that we can make in curriculum evaluation and empirical analyses. I don't think that one should be less importatnt than the other. I think, well I find in model states that I find the model very useful. Looking at the antecednets to an educational program, are those related to transactions and to the outcomes. The total interrelation there. If we can look at that in a logical way, we can also look at it in an empirical way, when we implement a program. We can set up various standards for judging performance, and we can then compare the obtained performance with the various standards, and we can use different groups to make judgments. So we might have curriculum evaluators make judgments. We might have experts in areas of special education make judgments. I think we need to utilize various interest groups in the community in evaluations of this type. Collecting priority data of this type in one project which Bob stake and Dennis , and I think could prove very fruitful. This type of analysis.

Mann: Julian?

Stanley: Well I want to say something a little oblique to this. It seems to me that in all these concerns - evaluating curricula or whatever we have to have a balance between being concerned with products and being concerned with process. We tend to be either too concerned with products in a black box sense, or too concerned with process in the sense that we don't worry about it. We have to have this instrumental relationship between process and product.

We don't want a curriculum for the sake of curriculum. It is just as we don't want to teach certain characteristics just for those characteristics sake even though outsiders might lack them. We want to be concerned with what this pupil is learning, how he is developing. That is the purpose of education. In so far as we focus too much on process - too much on product too, I think we get in trouble.

## REFERENCES

Adamson, G., & VanEtten, G. Zero reject model revisited: A workable alternative. Exceptional Children, 1972, 38, 735-738.

Apple, M.W. The adequacy of systems management procedures in education. Journal of Educational Research, 1972, 66, 10-18.

Baker, R.L. Curriculum evaluation. Review of Educational Research, 1969, 39, 399-358.

Berlak, H. Values, goals, public policy and educational evaluation. Review of Educational Research, 1970, 40, 261-278.

Caro, F.G. Issues in the evaluation of social programs. Review of Educational Research, 1971, 41, 87-114.

Cohen, D.K. Politics and research: Evaluation of social action programs in education. Review of Educational Research, 1970, 40, 213-238.

Frohreich, L.E. Costing programs for exceptional children: Dimensions and indices. Exceptional Children, 1973 39, 517-524.

Grotsky, J.N., Proger, B.B., Warkomsky, R.C., Hayden, D.L. * Murphy, W. Guidelines: PEP, Peer evaluation program: A systematic approach for evaluating educational programs. Harrisburg, Pa.: Bureau of Special and Compensatory Education, 1973. Available as ERIC document ED ******.

Guralnick, M.J. A research-service model for support of handicapped children. Exceptional Children, 1973, 39, 277-282.

Herriott, R.E., & Muse, D.N. Methodological issues in the study of school effects. In F.N. Kerlinger (Ed.) Review of Research in Education: 1. Itasca, Ill.: F.E. Peacock, 1973. Pp. 209-236.

Jones, R.L. Accountability in special education: Some problems. Exceptional Children, 1973, 39, 631-642.

Proger, B.B. Program evaluation: The model-building game. Journal of Learning Disabilities, 1970, 4, 292-306.

Roberts, F.K. National recognition for accreditation of residential schools for blind students. Exceptional Children, 1972, 38, 742-743.

Rosenshine, B. Evaluation of instruction. Review of Educational Research, 1970, 40, 279-300.

## References

Schutz, R.E. Methodological issues in curriculum research. Review of Educational Research, 1969, 39, 359-366.

Sjogren, D.D. Measurement techniques in evaluation. Review of Educational Research, 1970, 40, 301-320.

Stake, R.E. Objectives, priorities, and other judgment data. Review of Educational Research, 1970, 40, 181-212.

Welch, W.W. Curriculum evaluation. Review of Educational Research, 1969, 39, 429-443.

Westbury, I. Curriculum evaluation. Review of Educational Research, 1970, 40, 239-260.

Witmer, D.R. Cost studies in higher education. Review of Educational Research, 1972, 42, 99-127.

Yee, A.H. The limits of scientific-economic-technological approaches and the search for perspective in education: The case of performance contracting. Journal of Educational Research, 1972, 66, 19-29.

Yee, A.H. Every kid a winner: Accountability in education - A review. Journal of Educational Research, 1972, 65, 301-304.

# CHAPTER 9

## ACCOUNTABILITY: MEASURING TEACHER EFFECTIVENESS

Overview: Chapter 8 began discussions on accountability from the standpoint of evaluating a total program across teachers; the program was held accountable, not any individual teacher or administrator. However, an equally important aspect of the accountability movement is to evaluate the instructional effectiveness of any given teacher. There appear to be at least two major approaches to teacher accountability: (a) presage teacher variables, and (b) student behaviors. While the two approaches need not be mutually exclusive, the primitive state of research in this area has kept the two separate. In the presage approach, one looks at what the teacher brings to the instructional task both in terms of her previous experiences and her current, overt teaching acts. For example, one might examine her knowledge of the philosophical, sociological, and psychological foundations of education. Another facet to the presage approach is the activity that the teacher engages in during instruction; classroom observation schemes are often used in this regard. Standardized tests (such as the Minnesota Teacher Attitude Inventory) are often used to measure some aspects of the presage approach. Unfortunately, one still does not know what constitutes "good" teaching or the "good" teacher. Thus, the presage approach has been criticized in many circles. Thus, some researchers have turned to the student behavior approach, where one does not attempt to define a "good" teacher but instead assumes that a "good" teacher should produce a high rate of student learning in the cognitive, affective, and motor areas. To get a measure of how any given teacher affects student performance under controlled conditions, some have turned to microteaching. One devises a small, manageable unit of instruction or lesson plan, than asks the teacher under scrutiny to teach it as best she can with measureable goals in mind. The student performance on these goals is taken as the measure of the teacher's effectiveness. Regardless of which of the two approaches one favors, the area of teacher accountability can be quite frutiful, as well as volatile; certification, tenure, merit raises, screening by school districts for employment purposes, formative feedback in teacher training institutions, etc.

Mann: So let us move over to a topic that Jay's working on - the assessment of teacher effectiveness, this is a general issue but it certainly has implications for special education. So Jay could you lead off on the issue of teacher effectiveness?

Millman: Well, there are separate philosophies about how you measure teaching effectiveness. Perhaps one categorization of these philosophies was given by Paramitza some time ago when he referred to preshee variables, which are variables that the teacher brings in at her age and her training and the number of hours in special education and that sort of thing. These in a sense are indicaters of how effective a teacher she's going to be. However, I don't think many people take them very seriously at present. The two major competing philosophies concerning the assessment of teacher effectiveness I think can be summed up by various expressions. The one philosophy says it all, "It is not whether you win or lose the game it's how you play it type of thing." The group holding it would say to judge teaching effectiveness you go in with observation schedules that have been developed lately. You look and see how a teacher is teaching and whether she's using appropriate language..... Whether she shows more enthusiasm...Whether she has a high direct or indirect teaching ratio and so on. You look at the process and you make inferences from that as to how effective she is as a teacher. The other approach is the expression given by a Phillip 66 oil company, "It's performance that counts." or if you like something a little more religious, "by your proof yea shall know them." There the philosophy holds that you judge teaching effectiveness by the changes in and improvement of student behavior, skills, analogies, attitudes, interests.. Whatever you're criteria are. My own bias in the matter is that the goal of education is to change students; that's the whole purpose to change them in certain ways. In my scheme of things the more approrpaite way to get indicaters of teaching effectiveness is to look at what has happened to the children. Now the difficulty here is that to make comparisons among teachers as to their effectiveness as change agents you have to cut a lot of other intervening variables that can otherwise throw waves into the system. It's not fair to compare the teachers and their effectiveness if one teacher has an easier teaching task than a second teacher does. So it seems to me that if you want good data from research purposes on teaching effectiveness you need to provide these teachers in your studies comparable teaching tasks. By that I mean they should be teaching the same types of skills to groups of students who are as comparable as possible; perhaps randomly assigned to the teachers. That's the approach that I would suggest to use to get these kinds of measures. But the big problem in doing research on teacher effectiveness has been the lack of good criterion measurements and that is what we need find. Then when you have these, what I call, more purified measures of teaching effectiveness. Then it seems to me we can really go to work and find out what are the techniques the process variables, things the teacher might possess or actually do with a child that seems to be important and under what conditions.

me

Mann: Jay, do you place much faith in traditional teacher effectiveness?

Millman: Well, we can find out. I think when we get these good measures of teaching effectiveness then we can do the necessary correlations and see which tests are worthwhile and which ones aren't.

Mann: Is it just that criterion variables are not there at the present time, really? Not cleanly, anyway?

Millman: Well, that's my opinion. That is, I think the only most appropriate criterion variables are the effects the teacher is able to have on the pupil. However, you can't just look at the raw effects because of all these other things. I think the methadology is there now to get cleaner measures. Then we can look and see how about all these predicters. I'm less concerned right now about looking at scores on tests and other kinds of predicters because of my philosophy indicated this morning. I think good teaching is a science as well as a class of art. I think we can train teachers to be good teachers, but I don't think right now we have very solid evidence as to what she ought to be doing to be a good teacher. We have our opinions but we don't know for example about things like enthusiasm. When we look at these kinds of things in people behavior it turns out that enthusiasm or indirect to direct ratio turn out not to be related to change in pupils whereas from other things like general clarity of instruction variables tend to be related. I certainly don't think the data are all in. I don't think the story's complete but I think we can separate out now thse which are sort of folklore and sort of personal opinion kinds of things. Those practices which really make a difference in certain kinds of learning for certain kinds of kids.

Feldhusen: I guess, then, to be really sure that I understand your point, you were saying that you weren't much interested in tests of teacher effectiveness or didn't see this as very hopeful in helping us come to understand much better how to make a good teacher.

Millman: Well, I'm saying well I don't know. My guess is that something like the Minnesota test for example isn't going to correlate with the kinds of criterion measures and effectiveness I had in mind, but if it did correlate with it then I wouldn't nevertheless want to rule out immediately, that a teacher with negative scores could not be trained to be a good teacher, or a better teacher.

Proger: Well that's sort of a humanistic bias. We can't fault you for it, but it's neither here nor there as far as measurement and prediction are concerned.

Feldhusen: Well, it does turn out that sometimes, not in teacher training I don't think but in some other settings that perhaps Ryan's teacher characteristics do give us some hooks on, the criteria of teacher effectiveness. We may not be clear as to criteria of teacher effectiveness in any case but at least if the predicters were related to a decent criteria it does seem to me that one can get inferences from predicter variables that lead the way to things you could try for modification.

Millman: Let me say that for the record that there have been instituted, a series of U.S.O.E. supported teaching education efforts, I believe there's about nine of them in the country. Syracuse, University of Massachusetts, a number of these teacher training institutions have received a large amount of money and they all calim to be performance-based teacher education. I want to make very clear that they're not talking about the same thing that I'm talking about. When they say performance based they frequently mean teacher performance not sutdent performance and they're talking about when given a question of who John Dewey is the teacher will be able to know this and that. In other words they spell out very clearly the skills that the teacher ought to have, and many of these skills are practice skills. The teacher shall not talk with her back to the class and facing the chalk board. She will do this and have these kinds of skills and that's also performance based but it's performance on the part of the teacher.

When I talk about performance based and teacher effectiveness measurements I mean the performance of the student.

French: What kinds of criteria are you looking for?

Millman: That depends on the user and on the program. I mean for some cases I would think there might be some motor skill development, it could be cognitive development, it could be affective whatever you want.

Mann: Let me give you an example - so we have these three special learning disability classes and basically two instructural variables reading and arithmatic. We have hundreds of instructional objectives for each providing some sort of a ladder for the teacher to climb with the pupils. The children are well selected so that the groups have reasonably comparable degrees. I would think you could really identify the comparative achievement of teachers with the pupils as per their achievement of objectives.

Millman: If the children are comparable, if the instructional objectives are the same then it would seem to me a very fair way to compare the teaching effectiveness of your tasks. However, someone could come along and say but teaching is a lot more than just accomplishing go up the ladder in arithmetic and reading. Teaching is instilling a sinse of work on the part of the pupil, etc. And I would nod my head in agreement and say - if you believe that that is important then you need to see what is happening to their sense of work. But it seems to me to play the game fair the teacher who is being assessed ought to know the criteria on which she is being assessed. In other words what are the objectives. It would seem to me in your case it's very clear to the teacher that it is her job to help the kids up the ladder.

Mann: Bart?

Proger: I don't think I can go along too well with the teacher effectiveness. The current state of the research isn't encouraging. There are such a hugh number of characteristics, we haven't the slightest idea really of what an "effective teacher" is. You can use pupil performance, but I think that's a different bag again. We don't know if it's the technology of the instructional process itself that the teacher uses, her personality, her skills or whatever. I don't even see how you'd begin to attack the problem of what constitutes a good teacher.

Mann: Well, Jay is saying you establish clear criteria for your product, what a good product is, then you can get your predictors established.

Proger: I think you are implying the cause-and-effect relationships here. There may be correlations but I don't think they will tell us about teacher effectiveness.

Millman: Well, I couldn't disagree more. I don't know whether Bart and I ought to carry on a private little debate. I'm saying that you can ask what is it, what do you mean by being a good teacher. It is true people can give all kinds of answers. They say a good teacher is one who pays his union dues; a good teachers is one who will be congenial to colleagues on the floor, etc. But I would say many of these are either of secondary importance or are at best merely instrumental in achieving what I call the ultimate criteria of teacher effectiveness. Mainly to bring about changes in students in accordance with previously specified objectives. I think this is what education is all about. So I don't see the definitional problem as much of a hurdle. I think it's clear. A good teacher is one who can bring about desired changes on the part of the pupil. How do you find out if a teacher is good? You put her in a standardized situation where she has to bring about prespecified changes in pupils in the line of prespecified behavior. Now if there are objections, one could say, "Well, if the teacher can bring it about in a short period of a testing time, could she bring it out over the whole semester?" Could a teacher who can teach somebody arithmetic in one session really be able to teach him that and this. Any testing situation is only a sampling of all kinds of situations that might be possible and so on. But I think the good one is clear and I don't think it's a hopeless path. Maybe hopeless is to find out what caused her to be ineffective. But that's a whole other issue. We're addressing ourselves to how can we find out if this teacher is being effective.

Bracht: Jay, to what extent do you see this as managable, feasible. That we create standardized settings in which we can observe the performance?

Mann: We've done it already. We have two criteria - success in achieving instructional objectives in reading and arithmetic. Now you can certainly see where teachers stand in comparison to each other. If the same teachers want the same way again and again then I think we have a ranking of who is effective.

Bracht: But say we have student or graduates of a teacher education program and we want some idea: They're ready to go out for jobs, school districts are interviewing them, we want some information about thousands of teachers. Is this a managable thing?

Millman: We have a pilot study underway in which we are going to find out. We think it might be managable, in which the various school districts would use their own pupils to be taught about topics that the students don't know very much about because they are enrichment type topics. We think it is feasible. We don't know for sure. Talking psychometrically whether it's feasible, in a recent review of the generality of teaching effectiveness measured with less appropriate instruments, they still found that teaching does have a degree of generalized ability with correlations about .4, .5 when tested on different topics and so on, which gives me some hope that we can get reasonably good measures of teacher effectiveness. I think managementwise it is feasible.

Bracht: I'm concerned about another thing. You started your discussion by saying that we couldn't have a useful test unless we asked each teacher to perform in the same type setting. Students of the same ability or the same type classroom. That's what I see as perhaps the unmanagable thing. We have a thousand teachers and we want to get some idea of teaching effectiveness relative to one another or relative to some standard. We need to test them all in the same situation?

Millman: Well, a thousand is pretty high, but if you're talking about a staff of thirty then I think it's quite feasible. Remember, we're talking about very short instructional units. We're talking about 15 minutes of instruction in a unit and maybe three separate units. We're not talking about a great deal over a great time. You may say, "Can a teacher put something over in 15 minutes;" maybe she's a slow teacher or something. So far we've found that an increase in the amount of time allowed doesn't matter in the final results.

Proger: Could I ask a naive design question about this? You try to put the teachers into comparable situations, as far as the kids go, and what they're doing, the material to be conveyed to the kid, etc. and you look at pupil output as the ultimate criteria. Did the teacher do her job? Was she effective or not? In other words, try and dichotomize on pupil performance as per the unit, and then perhaps look backwards and say, "Alright, here I have the best performance and the porrest. I'm going to try and pull out a profile of teacher characteristics, amybe gotten through interaction analysis or some ovservational scheme that we're doing at the same time." Does this follow?

Millman: We're not studying the components of the teaching act that seem to be important in making for effective teachers. It could be done. All we're concerned with right now is devleoping psychometrically sound procedures for measuring teaching effectiveness. Once we have found those procedures then there are lots of questions to be answered. We've got some research underway which are looking at nine different administrative questions about the optimum way of administrating these tests. Now once we get what we think is a justifiable procedure then it seems to me that this good instrument could then be used for research on questions that we are raising.

Proger: What's going to become of the outcome? How are you going to describe effective teaching?

Mann: On the criterion variable.

Proger: No, you just don't say, "Well the kids did better or they did worse, and thus we have a good or poor teacher." It's too much of a black box; we'll never know why we found what we did. We'll never know if it was just a good crop of kids rather than a good teacher. Your trying to define what an effective teacher is. That's the ultimate output.

Millman: Well, I'm trying to give her a score. We can say Miss Jones ranks in the 93rd percentile in this test of teaching effectiveness.

Proger: But you don't know what she did, just that she's a good teacher.

Millman: That's right, and she's not told what to do. In other words she's not told, "Listen, you must teach in this kind of a style or that kind." She's given a job to do and we'll see.

Mann: That's a good philosophy, basically. Rosa?

Hagin: I just wanted to ask a question. You used the word teacher and teaching more or less interchangable. Are you concerned with teaching effectiveness or of the teacher's effectiveness?

Millman: We are measuring what happens to kids when a person trys to teach them something. Now that's what we've got, and we call what happens to kids a measure of teaching effectiveness.

Hagin: Teaching?

Millman: Well, it sounds a little less threatening than if we say teacher effectiveness.

Hagin: And more modifiable.

Hopkins: Well, I'm very glad to see someone taking this tactic toward teacher evaluation. I think one can judge beforehand really how much yield there is going to be. My own personal feeling is that ther's really a chance that something might come out of it. I'm not real clear on use, the particular use to which the results are going to be put: such as, teacher evaluation, teacher selection, or as a measure to study the instructional teaching process and find out which variables are related, or all of these.

Millman: I'd say all of the above. If we have good measures of teaching success, then we could use the measures to do research on the instructional topics to identify what are the variables that relate to teaching success. If we have good measures of how successful teachers can be in teaching, we can use those measures to help us in the selection of teachers or to evaluate them. Further, if we have good measures of how successful teachers can be, we could use the tasks in certification. Certainly, I think that giving the teachers the opportunity to try and teach lessons and giving them suggestions as to that thing they might be able to do to improve their ability to get material across to the student, it would also be very useful in training both at the preservice and the inservice level. So I see the usefulness of having good measures of teaching effectiveness extending into preservice/inservice education, into teacher certification, into teacher selection, and into research.

## REFERENCES

Aleamoni, L.M., & Yimer, M. An investigation of the relationship between colleague rating, student rating, research productivity, and academic rank in rating instructional effectiveness. *Journal of Educational Psychology*, 1973, 64, 274-277.

Blackwell, R.B. Study of effective and ineffective teachers of the trainable mentally retarded. *Exceptional Children*, 1972, 39, 139-143.

Bledsoe, J.C., Brown, I.D., & Strickland, A.D. Factors related to pupil observation reports of teachers and attitudes toward their teacher. *Journal of Educational Research*, 1971, 65, 119-126.

Borg, W.R. The minicourse as a vehicle for changing teacher behavior. *Journal of Educational Psychology*, 1972, 63, 572-579.

Costin, F., Greenough, W.T., & Menges, R.J. Student ratings of college teaching: Rehability, validity, and usefulness. *Review of Educational Research*, 1971, 41, 511-535.

Davison, D.C. Perceived reward value of teacher reinforcement and attitude toward teacher: An application of Newcomb's balance theory. *Journal of Educational Psychology*, 1972, 63, 418-422.

Dobson, J. Predicting and evaluating student teacher behavior. *Exceptional Children*, 1972, 39, 29-35.

Emmer, E.T. Good, T.L., & Oakland, T.D. Effect of feedback expectancy on choice of teaching styles. *Journal of Educational Psychology*, 1971, 62, 451-455.

Finkbeiner, C.T., Lathrop, J.S., & Schuerger, J.M. Course and instructor evaluation: Some dimensions of a questionnaire. *Journal of Educational Psychology*, 1973, 64, 159-163.

Gardner, J.M. Selection of nonprofessionals for behavior modification programs. *American Journal of Mental Deficiency*, 1972, 76, 680-685.

Koran, J.J., Jr., Koran, M.L., & McDonald, F.J. Effects of different sources of positive and negative information on observational learning of a teaching skill. *Journal of Educational Psychology*, 1972, 63, 405-410.

Levinthal, C.F., Lansky, L.M., & Andrews, O.E. Student evaluations of teacher behaviors as estimations of real-ideal discrepancies: A critique of teacher rating methods. *Journal of Educational Psychology*, 1971, 62, 104-109.

## References

McNeil, J.D., & Popham, W.J. The assessment of teacher competence. In R.M.W. Travers (Ed.) Second handbook of research on teaching. Chicago: Rand McNally, 1973. Pp. 218-244.

Miller, M.T. Instructor attitudes toward, and their use of, student ratings of teachers. Journal of Educational Psychology, 1971, 62, 235-239.

Prien, E.P. Measurement of performance effectiveness of hospital attendants. American Journal of Mental Deficiency, 1972, 77, 21-25.

Quirk, T.J., Witten, B.J., & Weinberg, S.F. Review of studies of the concurrent and predictive validity of the National Teacher Examinations. Review of Educational Research, 1973, 43, 89-113.

Rochford, T., & Brennan, R. A performance criteria approach to teacher preparation. Exceptional Children, 1972, 38. 635-639.

Salomon, G., & McDonald, F.J. Pretest and posttest reactions to self-viewing one's teaching performance on video tape. Journal of Educational Psychology, 1970, 61, 280-286.

Tolor, A. Evaluation of perceived teacher effectiveness. Journal of Educational Psychology, 1973, 64, 98-104.

Wagner, A.C. Changing teaching behavior: A comparison of microteaching and cognitive discrimination training. Journal of Educational Psychology, 1973, 64, 299-305.

## CHAPTER 10

## CONSIDERATIONS IN STANDARDIZED TESTING

Overview: Standardized tests are used more by both teachers and measurement specialists in regular education, in contrast to special education where psychologists are the primary consumers of tests. In both areas there is a great deal of misunderstanding surrounding the uses and abuses of standardized tests. Appropriate test selection has plagued everyone in education. Consumers are often unaware that standardized tests include not only paper-and-pencil measures but also performance tests, unobtrusive measures, and the like. Test consumers also often become overcome by the technical aspects of testing and fail to bear in mind commonsense guidelines. A second major source of confusion centers around the concepts of validity and reliability. There are a variety of issues involved in the adequate development and release procedures of new tests.

**Mann:** O.K. Do we want to bite on validity now?

**Proger:** Unbelievably little validation evidence exists in the tests that are released for special education purposes no matter how big a publisher or how well known the authors might be on the test. John, in his paper on juvenile delinquency said something about reinventing the wheel, that a lot of the researchers, when they seek tests appropriate for predicting delinquency or whatever try to find something newer and better and don't really look at the existing instruments that might suffice, and he brings up this point about reinventing the wheel. This certainly holds in special education test research. Everybody shoots down the other guy's instrument and says, "We'll go out and build a new one, a better one." And then we find out, lo and behold, there's no supportive validation evidence for the new one. I think Joe mentioned this too. Isn't it time we sit down and look at the existing instruments and plan out a systematic program of validation research? As Joe said at the convention that some test inventor said he's finally glad his instrument is out now after it was demonstrated to be pretty sound in certain areas.

**Millman:** Well, one could talk on and on about test validity and I probably will, but before I do let me ask a couple of questions. What are the kinds of things you had in mind Bart? I think it would be rather helpful to me to hear a bit of discussion on what people are uncomfortable about before I for one start launching my tirade concerning validity.

**Proger:** In many cases of new test reviews there is absolutely no validation or what little evidence they might give is misleading to the general test-consuming public. A number of us have become very interested in the idea of convergent and descriminant validity of a test as perhaps being the Cadillac of validity research designs and all we ever get is convergent validity. A test is found to relate highly with other existing instruments in the same area, yet we don't know how it relates with totally different traits that it should not relate very highly with. The evidence on convergent validation itself is very skimpy. That's what shocks me. I would assume when somebody goes out and spends who knows how much money and gathers thousands of bodies for standardization purposes across the country that we could at least get a few coefficients of convergent construct validities.

**Hagin:** I can give a specific example of what has troubled us. I think the APA doesn't cover this in their recommendations because these tests come out as experimental editions. So the test standards don't apply and obviously one must experiment. But what troubles me is that we continue using experimental tests and apply their results as if they are established instruments. There isn't anything better and we have difficulty finding criteria for validation. The things that are on the market are validated against existing measures which are also questionable, and we can't blame the test makers. Validation is a very expensive proposition. I know something of what has gone into the WISC and when people now ask Dr. Wechsler about validating his tests he says, "Spend your own money." To be specific about one of our problems, Wepman's Auditory Discrimination Test, is the modest little test which has been out for about ten years, with very little information about how it was constructed and how it was standardized. Now we have found using it that the results are not only unreliable but also that we aren't sure what the children respond to on it. They must listen to two words and make a judgment of same and different. Reportedly this can be used with 5 and 6 year olds. We find that our 5 and 6 year olds don't know same and different and respond to all kinds of non-intended things, or just lack understanding of the concept involved. Yet this test is used repeatedly as a measure of auditory discrimination.

**Millman:** OK, I want my remarks to be relevant. Would you talk just a bit about the Metropolitan Readiness Tests which you used, Rosa, and which is a test that most people would agree much effort has gone into. A well know, widely used test. Now you also found something unusual with that test with a special group.

**Rosa:** We found with our group, the mean IQ, the first grade-performance IQ, was 99. something with a standard deviation of 15. When we gave the Metropolitan Readiness Test we found that it did not select the children who at the end of the first grade were found to have learning disabilities. It selected the children with low IQ's. Then we went back and did an item analysis of the content and we found that the content of the test was inapprorpaite to the children. There are things like collie dogs that these children did not recognize. The teachers also felt that many of the children just did not know how to respond just how to put answers down on the test; it was beyond what they could do at that age. I think it is a naive assumption to take the scores with our group and use it for predictive purposes.

**French:** Well, one other problem and that would be to me the flyers we get from some publishers describing very high almost unvelievably high correlations in predictive studies. But if you search to the original source you find that it might have been on 20 cerebral palsy children, who ages span from 5 to 20 misleading correlations.

**Feldhusen:** Well, test manuals are known to be notorious kind to questionable data about such things; whereas, the Mental Measurement Yearbook stands at the opposite end of the spectrum, being usually severe. Now, is it the case that much of the measurement instruments in special education somehow or other are not in the standard test literature, and therefore do not come under review?

**Mann:** I've done some reviewing for Buros. The problem is that the present proliferation of instruments makes it all but impossible to survey the field. Buros expressed to me his frustration in trying to manage this unending flow of tests and this included special education.

**Proger:** How would the APA standards help offset this uncontrollable state?

**Mann:** I think invariably you'd knock a lot of tests off the market that might clean up the situation.

**French:** There should be a fair practices marketing act of some kind which would reate the test and each of the factors as to whether or not they've fulfilled the most desirable characteristic desirable and so on.

**Mann:** That's not a bad idea in fact for the journal. We might even try that out.

**French:** Well, it takes a lot of work. Somebody out at UCLA tried to do that with preschool tests and with elementary school tests and they had grief, but I think they tried perhaps too many tests with too few people. It needs to be very critical and careful evaluation.

Millman: General understanding on this sort of thing is difficult because of course different people might want to do idfferent things with tests. One may want to make different kinds of inferences from test results. A test might permit valid inferences of one type and not valid inferences of another. This is the old story that a test does not have a single validity to it, but rather might be valid for some purposes and not for other purposes. I would think it would behoove the user of any instrument to ask himself, what is it that I want to use these scores for? What kind of inferences do I want to make on the basis of my scores? And then to say, are these inferences accurate using these scores? Now to give you some for instances. If the inference that you want to make is to use the test score in order to infer how a person is going to do on some later performance, that is a predictive type of inference, then what you want to do is see literature data, if you will, on what are the correlations between the test scores and these later performances on kinds of kids that you're going to be making these predictions about. That's straightforward enough. Now, however, we're not always interested in predictions. Sometimes more often interested in just ascertaining some skill or trait of the individual like lets say auditory discrimination. Now if you purpose is to make an inference about how well a person can discriminate certain kinds of auditory sound then the appropriate thing seems to me to get clear in your own mind what kinds of sound are you interested in knowing if you can discriminate. You have to look at the test, do a lot more eye balling, see what kinds of tasks the child is being asked to perform, On the basis of how he performs you're going to make this inference. Now if you see that he's got to carry out a task on the test item where he has to know the concept of same and different and you feel that he could discriminate then obviously that test is not going to be valid for your purpose; that is the inferences you make on that test. I guess what i'm saying is that I'm not sure we can given a single rating for a given test. I think that perhaps that approach is not going to be the salvation. I'm not bojecting to more information. It's just that I think that every user of the test has to be clear in his own mind what inferences he wants to make from the scores he gets and if those inferences will then lead to what kind ofinformation you want to find. Freqnently, one must go to the test itself and look at the collie dog in order to understand the issues and problems. But I don't think for many of our purposes we need a great deal of empirical kinds of studies and so on that we sometimes think we want. I think sometimes homemade instruments where you know exactly the kind of thing that you want to make inferences about perhaps might well be the best answer.

Proger: I'm totally disillusioned. I agree with the idea that we have many separate validity purposes to think about in looking at a test. I guess what disturbs me about your comments Jay is that just by looking at content validity of the test alone, sort of an eye balling the whole test, I don't think that's going to help a test consumer that much, because most of them don't know what to look for, unless they've had a lot of experience with it or gone through a lot of problems with it. But somehow in my mind construct validity is of utmost importance in any instrument no matter what the primary practical use might be: assessment, or prediction, diagnosis, screening, and so forth.

Millman: What I was saying, are constructs really, you know when you want constructs do you want a test that measures some sort of construct like anxiety about? Well then obviously you've got to collect the kind of data one usually thinks of when he does construct validity -- discriminant convergent validity information - I'm all for that. Looking at nomological networks, I'm all for that. But I think often you're also interested in areas in which the person has certain ability, skills, certain understandings, can he do this, can he not do that? There I think your definition of what is the universal item the one that is relevant foryour interests and should dictate the nature of the items you concern yourself with on the test. In other words I'm saying that maybe you should worry a lot more about the operationalizing of some of these things that we're interested in.

Mann: Glenn?

Bracht: I think Rosa's illustration is a gook illustrate of issues to content validity. Here is perhaps a good test which meets our general needs. Maybe certain items like the collie dog are not good items in that test. But maybe for the most part we do have a good test with good content validity. But the population is not able to understand the procedure of sameness and differentness and so the test really does not measure what we want it to measure and so construct validity is lacking in that particular subpopulation, although content validity may be there.

Millman: I feel it's the same difference with the auditory discrimination test discussed. The question is do we have predicted validity for the test for this subpopulation, and there the inference that one wants to make from the test is: can it predict which kids are going to have learning problems a year later? Obviously there, what you want is evidence on what kind of thin. You don't eyeball a test that is supposed to make a prediction; you see if it can make the prediction.

French: I think that's the key thing. When we construct tests locally or items locally or even use published things that we determine how well they work. Years ago I was talking to some people who were using the DAT to predict success in a vocational education course. For one school the correllation was fantastically high and another school in the same state the correlation was negligible or zero. We never did discover what the problem was, but one school quit using the test when they discovered it was no help in prediction and the other school continued. I think that when we use tests construct that have some success that we sometimes get over confident, and feel we don't need to constantly evaluate what we're doing.

Proger: Are you indicating, Jay, that you don't particularly get worked up at the lack of empirical evidence on a lot of these things? I'm rather surprised.

Millman: Well, I had two things in mind when I said this empirical business. One relates to the fact that for many purposes where you are interested in measuring say, certain skills or abilities or something, you in a sense are looking at the universe of possible behaviors from which you want to sample. Here I'm not sure you need a lot of empirical support. Or a Herman Mills kind of test; you define the universe of possible behaviors and you sample from them. So that was the one. But even for the second, a prediction situation, for example, well! I know the DAT you mentioned. The manual is full of all kinds of correlations showing the relationships between DAT scores and achievements in this and that. For my purposes in predicting, using the DAT prediction those correlations would not be very helpful. So what I'm saying is that just because there's a lot of empirical data the nature of the inferences I might want to make might mean that those empirical data are not very relevant for me. I'm also saying that we don't need more empirical correlations of tests with this and that and so on. But rather what we need is certain information that is relevant to the inferences that you want to make from those scores from those tests.

**Feldhusen:** I agree with everything you say, but I do have a fear that consumers of the information in special education research might see your recommendations as discouraging any sophisticated test evaluation.

**Bracht:** It seems to me that in research the primary purpose for using tests is to provide a description of the subjects behavior for I think we need to carry out validation studies which will provide evidence that the test is providing accurate descriptions of the subjects behavior. I don't think validation studies are ever completed on a given test. There is always another study that can be done, or another situation in which the test is used for which we don't have enough evidence that the descriptions are accurate. So it's always a continuing process and we can never be satisfied with what we presently know. I personally find Cronbach's chapter in the recent education and measurement book very helpful in this respect. Among the things he recommends when using tests includes a clear description of the populations. So in doing a validation study we first need to clearly define the populations for which that validation study has been performed. Perhaps when I use the test my subjects may correspond to that population which we have valid data, but then again perhaps they do not correspond to it so I really con't use the validity data from the other studies to make decisions about the validity for the test I want to use. Another thing Cronbach describes is the conditions under which the test is given; the procedures that are followed. I'm reminded here of a study that was published in the American Educational Research Journal, regarding the use of creativity tests. I forget the authors now but what they discovered is that when giving creativity tests to some students who were in an environment that was somewhat exciting and interesting that the students did not perform very well on the creativity test. The creativity tests were kind of a let down from the environment that they were just taken from. Whereas, other students who had been in a somewhat boring environment were quite stimulated by the creativity tests. So I think that the context in which we give tests is a very important consideration. I think the important thing as per content validity is that we have a very clear definition of the content domain when we develop tests or if we are looking at someone elses test that we try to get from the test author if it is not available, a clear description of the content domain. Personally, I'd like to use some type of grid, laying out test specifications regarding content and maybe the type of behaviors or levels of outcome that are to be tested. But there are other procedures for defining the content domain of the test. Cronback also talks about educational importance, and this is something that the test consumer must be concerned about. What is it that I really want to measure in this research setting? The tests that I select must then correspond to what I want to measure. Quite often we look at educational importance in a post hoc sense. After we've done a study and we didn't find expected results but yet we have that intuitive feeling that something really did occur, we go to the test and say well the test really doesn't measure what I wanted it to measure. I find researchers seldom doing this sort of thing before they do the study. So I would encourage that type of behavior. And then construct validation is the 5th thing that he talks about.

Mann: How about the convergent divergent approach as involved in construct validation. Bart keeps calling it the Cadillac of validation procedures.

Bracht: I don't know; I think it depends upon the test. Well I think a variety of things are important and it's a matter of relative emphasis. In one situation content validity may be of primary importance, as Jay has indicated. I think it just depends on what we're measuring and what we're using the test for. I think in many cases construct validity is the thing that we must give the most attention to. It's probably the thing that needs the most investigation. So it just depends upon the use of the test and the type of test.

Mann: During one of our breaks, Jay, I think a number of us here expressed dismay at the eye balling approach to validity recommended by one as eminent as you in psychometrics.

Millman: It depends on the purpose.

French: It depends on whose eyeballs.

Millman: People ought to look at the tests at the item level and really see what the test is like. Now for some purposes you can't run away from the need of empirical data validation. If the purpose is to measure some sort of hypothetical abstract trait, some construct, then that's very tricky business and you need all the information you can to see whether it's holding up for the kinds of people you would want to use it for. The Campbell-Fiske approach requires you to get information not only on the degree to which your test can correlate highly with things it's supposed to but also the extent to which you can discriminate certain types of ... that is correlated relatively high on some things and low on others. For construct validity the discriminant convergent divergent approach is perhaps the Cadillac. I think it provides very good kinds of information. But I don't think it's very useful for numerour other purposes in measurement that aren't concerned about how to measure a construct.

Feldhusen: I suppose that a very important issue that we haven't brought up, I don't think so far, that is extremely relevant. I first became aware in 1963 in a paper that Bob Glazer gave about the critical question of the test being sensitive to treatment effects in an experimental study. Now all of you know, of course, what many different directions that issue has taken as it blended into the norm versus criterion-referenced testing debate. But the later development of the norm versus criterion-reference testing issue does not always address itself specifically to the issue. Bob, did make the point: A test developed on norm records is designed to be maximally sensitive to individual differences and that is not only something that could put it out of the ball park in an experimental study but could even make the test troublesome because of the large amount of individual difference variance that is introduced by its normative construction. His basic point was that it is essential or at least extremely desirable that there is to be some evidence ( and I think this could be interpreted as validity evidence), that the test is sensitive to and would respond to treatment effects of some kind or other. That particular kind of validity I guess we still would look awfully long and hard to find for anybody's test. We can get all kinds of evidence that the test discriminated among people along a set of individual difference variables but anything to tell us that it might indeed be sensitive to change from a precondition to a postcondition after treatment that is another thing. It is a very big problem for most of the tests that we're using, I suppose 95%. And even if we develop a test for a particular study I doubt that we engage in procedures that would investigate the sensitivity of the test to change, as a result of any kind of treatment. I raise this as a problem and I think a critical one yet with maybe a solution's coming from the norm reference versus criterion-reference debate within test development.

Mann: Joe?

French: I'd like to speak for the test author about how difficult it is to obtain the kind of data that is desirable. Maybe a research and development center approach will be the only way to go to secure appropriate funding for validation. I think it's become increasingly difficult to get data on the basis of good will. Ten years ago I was trying to standardize an individual test for kids 4 through 8, and I was able to ride around the country and get volunteers from a very carefully controlled sampling procudure. I got cooperation in Oklahoma, Havana, Florida and a number of places where it was appropriate to gather data for sampling. I could write to people and get enough people to volunteer to standardize my test without very much in the way of money being spent. But as I listen to people today it appears increasingly difficult to get normative data on tests. School systems what to be paid. The teachers unions are against allowing the teachers to administer such tests and so on. As this goes on it's going to become an increasingly difficult thing to gather such data. I helped identify subjects for Wechsler's edition of the WPPSI and it was very very difficult to get people, even when paid $15.00 per examination, to go through the procedure of locating the subjects properly and giving the test properly and getting the data. It's becoming very expensive.

Mann: I think you're perfectly right I find staff's extremely reluctant to depart from their sole preceptions. They feel they're veing put upon and even when they cooperate in a test by the results may be tainted.

French: Sometimes it's superfically done when they don't want to do it.

Feldhusen: I guess in part that is because the world of researchers seem to have been hooked so completely on an ocean of paper and pencil testing and to be unable to conceptualize in any other way of measuring that we have just vastly over proliferated paper and pencil type testing and created resistances. I think this would be a good admonition as per Jay's suggestion to look closely at purposes of your assessment or what you're trying to measure. It would seem to me that more often people ought to be involved in something other than what is involved in paper and pencil measurement. This is a standing and I think s solid criticism of our research in education researching. Quite in contrast I'm impressed that my colleagues in the psychology department seem to be totally uninterested, except for a few people in clinical psychology, in paper and pencil assessment. They look to other forms of mechanization of measurement that just don't depend on pencil and paper.

Hopkins: I find it very difficult to respond to generalities we're getting here. Just like paper and pencil test, I think that's the very best way of measuring some things. If I'm trying to measure personality I think it's a very poor way, but If I'm trying to measure addition and all that and whether a kid can read or not I don't know how you're going to get away from paper. I guess you could, but it's not going to be a meaning ful improvement using a TV screen or whatever. So, of course, paper and pencil tests, I think, are very poor choices for measuring some kinds of variables and very good for other kinds of variables. The medium whether it's effectively used or not is something else. Of course, your individual intelligence tests don't rely that much on paper or pencil in most cases. So that one comment of mine. Another is about the response to treatment of various test measures. I think again in the educational area we know that they are very sensitive to something because second graders do a lot better than first graders and sixth graders do better than second graders. So I think we do have evidence of treatment effects whether it be maturation or education. Another isolated comment, the convergent divergent method business, you know I think that's a fine component of construct validation. But in predicting validity and content validity I don't think it's too important. Also I think it can be misleading if reliabilities are not taken into account. Creativity tests, everyone knows they didn't correlate rather highly with conventional intelligence tests so people took that as evidence that they were measuring different things. But were they too unreliable to correlate highly? And among themselves creativity measures we know, are not going to correlate that highly because of their modest reliability in most cases. And they may correlate as highly with conventional intelligence tests as they do with another creativity test. Because the conventional intelligence test is more reliable. So all of this doesn't really come to grip with the commonality issue. Well again that's a very scattered disjointed comment but I find some difficulty in responding to the more general issues here. I've got a number of such isolated disjointed comments.

Mann: Want to continue with them?

**Hopkins:** I think that this issue of norms, of whether different groups are represented in norms I think that sometimes is irrelevant too. If you're comparing some person with some group, I don't think that that person needs to be represented in some way in that group. I don't think you need to have blind people in the WISC standardization to see how they compare with sighted people in the WISC ranking. I can't get keyed up about the fact that many retardates weren't represented in the Wechsler standardization.

**French:** Isn't it important to try the test with retardates and then determine the predictive validity of the test with that particular sampling?

**Hopkins:** I couldn't agree more with the point that I think Glenn spoke of. You certainly con't generalize from any sort of validation group to a population different from that on which it was attained. Like in inner city predictive study of the Metropolitan. I also think it's a little unfair to expect the test publisher to produce that kind of data. I think it is the professions who should feel some obligation to see if the Metropolitan has predictive validity for the innercity black.

**French:** Well, then I very much agree with you and would like to ask you, how can we get this data available to the professions? Suppose a test does meet minimum criteria for publication and distribution, perhaps as an experimental edition, what can the profession do to assimilate the various very small validity studies that often wouldn't be published in existing journals but would contribute evidence to how the test fared with a particular subject?

**Proger:** You could have a clearing house where the stuff would come into, feed into, and be published in a manual at periodic intervals.

**French:** Well that's very difficult to do. A lot of people have written to me in the early days of publication of the PTI and asked for validation studies, and they were going to use it on a particular sample. So I'd write back and way well no one's used it yet except what appeared in the manual. I'd really like to have the kind of data that you generate. I received almost nothing, a masters thesis here and there. I'd write them two years later and ask them for information and again got very very little feedback.

Feldhusen: We could hope that the ETS ERIC would help here. Now I don't know what precisely all the plans are but ETS ERIC is to serve the business in the field of testing, and I presume that it can become a repository of such information as you are talking about here.

French: I guess what we need is a reinforcement schedule for people who gather that kind of data to get them to submit it. There are an awful lot of people who do research with children who don't send things to be published and where work goes unnoticed. I'm thinking about the person in the school district who gathers some kind of validity data about a particular instrument. There are not as many Brownie points for him to publish something as there is for a college professor. There isn't a reward system that gives him reason to send it to ERIC. It's more work if he send it to ERIC and there's no value derived from it for him.

Millman: May I ask what kind of data is usually available from these kinds of studies? I mean aren't the test items usually used in stituaiton where the data really aren't going to be very useful for the next user?

Mann: We could, of course, establish proper ways that they could gather this data. Many school systems have an enormous amount of poorly gathered, poorly established data. But they could be made a real contribution.

Millman: What I want to come back to is are we going to be much better off if every time a test is used, in either kind of a way are we going to be much better off if the data are assembled so that the user who is considering the use of that test will now have access to all the data that was gathered about it? I question that; for example, I'm an author of a test called Cornel Test of Critical Thinking, co-author, and we've gathered information about our test as it's been used in Masters theses and so on. I'm not sure that those data are going to be helpful particularly to future users in deciding whether or not our test is relevant for them.

Proger: Are you saying in extremes what's the point of validating a test essentially?

Millman: Some information is important and some is not. I guess what I'm saying is that what we need necessarily is not more information but information that is relevant to the certain purposes that we want to use the test for.

Bracht: If the test was used in a study only as a dependent variable and there's no other data collected then that information isn't very useful but if there is other test data where we can look at maybe convergent or discriminate validity then I think yes we would have some useful information.

Feldhusen: We persist in talking about tests, that have a relatively high level of developmental status, are as standardized and thelike and for which, if nothing else, there is title validity and careful development and what have you. At the opposite end of the spectrum, I want to ask is how much is the field of special education beset by the problem in the research literature of what I guess are best called fugitive tests, that is, the research report tells us that the researcher devleoped an attitude instrument to assess the attitude of the subjects toward something. Then he goes on and tells us the results of the study. A good illustration is the field of self concept research widely documented by Piers Harris, Tennicy and a couple of others in her book. Every researcher developed his own self coneept instrument. Partly because you could never get ahold of anyone elses instrument or reconcile results. Even though the researcher reports and you publish that he had an instrument you write and ask for it but you don't get it. I've been through this trial. Are fugitive instruments a problem in general education?

French: Well even if we go with Jay's point of that in many instances it's desirable to construct items that will relate with a particular thing. It's a lot easier to go to somebody elses's work and add a few items to it or to modify something that somebody else has already done, that to start from scratch and this has often been done in special education. There is an enormous variety of new tests generated. Reports are borrowed or adapted from the standard in the field. I'd like to see pools of items make available for use in different areas.

Millman: There are some collections of old tests in certain areas. I mean Buros is not the only place one can go to see where there are tests on certain topics. There are now, a number of books out that indicate tests, and I believe, Ken, in your institution they're also trying to get another collection Of I think in the achievement area out.

Hopkins: There was a project for measurement instrument exchange where they were trying to colelct them in a reservior, not to evaluate atall. Fugitive instruments to do exactly what Joe suggests. So that people who are developing a measure in this area will have something to work from. Modified not necessarily to use unchanged. It would save a lot of developmental work.

Feldhusen: However, if you are just collating by the title! Beyond that you have to group the content somehow to be able to see the thing possibly. As an eidtor yes, what do you do about an aritcle with a fugitive test course in test items in a table or appendix. We ask the others to give us the items or the breakdown of the test so the reader has some comprehension of what he's talking about. That would be a good notion. It is otherwise not untypical to illustrate a few items in other journals. Again the journal could have a rule: the rule could well be a deposit of that test with ADI or something or other, because it does turn out to be almost impossible as you know to get it form the author himself onve his publication is out.

Proger: Along the line of compendium of unpublished personality instruments, (talk about really far out fugitive instruments!) I think one of the most useful collections, not that old either, is the ASCD monograph. That has a lot of these in it. There are other similar efforts. Joe, was mentioning the tremendous expense now a days of producing validation studies and many standardization efforts. Rosa, too discussed with me her concern about "experimental" version of tests being readily applied as tried and true on the market. Cuoldn't these be something of a resolution here? Couldn't there be a policy with any new test remain experimental and not be sold until the test makers spend their own dough and their own time in research labs and anywhere else as properly standardized. Until somebody or a board of experts decide it is mature enough to be sold. I still believe in getting plenty of data. Any type of data and any type of validity is tremendously valuable to me, and I can't see how one can say otherwise, Jay. Then somebody can finally say, "OK now, it's perhaps ready. This is a test that's somewhat acceptable and no longer experimental." They'll still refine it, but at least it's on its way.

Rosa: The special education field is utterly naive and eager to be taken by the test that comes someplace between the fugitive instruments and what appears in experimental editions. As you know in the field of learning disabilities everybody has a little screening instrument and they set up their own publishing company in some cases and sell it and the field is eager to buy it and use it and not to experiment with it even in very simple fashion. I don't know what the solution is. One possibility might be that experimenters establish such good relationships with the field that it will be easier to get the kind of data that you said is increasingly more difficult to get. Schools feel that they have been used by researchers. Maybe they have; somehow it must be amended so that schools feel they get some feedback and can use the feedback. Because as we know education goes on doing it's own thing irrespective of what the findings are much of the time. Jean Chall made this point over and over again in her study of first grade reading, and yet was severely criticized for it. I think the researchers have something to offer in exchange; some feedback to the educational process. But I also think we have to do a better job in teaching teachers about using research.

Mann: Would you have some more of those random comments that were so ehlpful, Ken?

Rosa: Can I ask a random question about one of his random comments? You're feeling that it's not important, the inclusion of subgroups in the standardization clinical. I think this is a very sophisticated response; however, there are many groups who are angry and emotional about, for instance, ethnic groups being represented. Can you talk of this?

Hopkins: Well I think it is important to know about the groups against which you're making some kind of comparison. I think comparisons are most meaningful if there are representative groups. Representative of some population, like six year old kids. I think all kinds of people should be represented in proportion, in the same proportion they are of the population let's say since there are 200 subjects at each age level on the WISC maybe 2 or 3% of below 76 or 5 retarded kids that should be in the standardization population. I certainly think those norm groups should be as representative as possible. I think they're the most meaningful, but I do think, however, in the Wechsler Bellevue where there were no blacks or something like that as I recall, I don't think that negated the value of the norm group. I just think the norm would have been more valuable if they had represented all the people.

Mann: Let me phrase that a little more concretely perhaps because we've all raised this issue . Should there be a test of black intelligence for this subgroup which would be fair to them? Of course, this would make it hard for many whites in terms of the terminology used. Should there be an achievement test for the retarded? Which some people have claimed. It seems like a contradiction in terms and perhaps what's the point of it if you want to see how people move as per general criterion. But there are people who raise these issues.

French: Well an achievement test for the retarded, I think there's some value too in providing enough items that encable you to really study the behavior of the individual. On some achievement tests there just aren't enough items for the retarded to respond to.

Mann: No one argues we need more sensitive instruments that allows us to make these judgments. So that's perhaps the answer but not separate tests but... How do you feel about that?

Hopkins: Well if I understand the question, I agree with it completely. On most standardized achievement tests there is an insufficient functional floor and anyone who will lose faith very quickly, if using a standardized achievement test, let's say he gives the elementary achievement battery to 12 year old retarded kids because they'll come out looking pretty good if they just attempt all the items. So chance scores on standardized tests if ready can easily come out to be in the 25th percentile or the 30th percentile which looks pretty good, particularly when the kid can't read. I certainly hope people in special education do not administer good will standardized tests by functional levels so they might give the primary battery to the 12 year old youngster. They have to go by age for a meaningful assessment if the kids aren't up to grade level.

French: Often it's necessary to give the test in groups of two or three so that the administrator knows whether or not the kids are randomly marking the page or whether they seem to have some comprehension of what's going on.

Mann: There is the other side of the story. The youngster really has learned some functional arithmetic and reading but faced with the distractions and confusions of a printed test page he just can't produce a response which reflects what he has learned. This is an opportunity to chance scoring which gives him a high score. Other comments?

Bracht: One comment is that my experience indicates that researchers are sometimes too willing to use a test by it's title. I think we have illustrations of tests with similar titles or maybe one sentence or two sentence descriptions which indicate they measure the same thing. For example, a self-concept or something of that sort. Yet really the person who developed the test had a much different definition of self-concept from the different persons who developed his scale. I think that test developers are going to have to provide some clear definition of their scales, so that users have some basis of judging whether scales with similar titles really do measure the same thing. Both short of that I think at least test authors can provide a very comprehensive description of the domain in which the items were drawn. Whether it's from some sort of a grid or a set of objectives or what it may be but these one sentence, two sentence descriptions of what a test measures are highly inadequate.

Mann: Larny?

Cross: I guess I'll take a chance here. Again with respect to test prediction, which I guess Julian Stanley will answer to, I guess Darlingtons Spoketan, and Cleary and so forth spoke of the issue. There's a lot of clouds in there as far as whether it is useful to use the same test to predict for blacks. In fact I think some of the studies indicated that using the common regression, slope the SAT will over predict for blacks, contrary to the popular opinion it will under predict. But in answer to Rosa's question about whether or not the blacks should be included in the norming population, It seems to me it's quite a different question as opposed to retarded where we have the problem of the descrimination of a test at a low level.

Here if a kid has nothing right at a certain grade level, lets say he takes it at a sixth grade level, he has no items right he's still going to get a grade point average of like 2.2 or something of that sort. He knows nothing but-- so that's a problem. But lets suppose it is the case that for one reason or another the subpopulation be it black, Chiccano, or whatever, is 5 points less or if it's Oriental if their 5 points above, you still have tremendous group variability. In terms of the total population it's a relatively small percentage, that is, maybe 10% of the population in black, maybe it's more that 20%, I don't know certainly with Oriental it's less than 5% in our country. To include them in the norming group I wonder if your using a scaled scores for norming of planning as you incorporate this different group if it's really going to have that much effect over your scaled scores after you're done norming. In other words they have sikilar heterogeneity, the norming may be shifted slightly either up or down but when it's a small percentage of the population I wonder what the practical effect is going to be. I think it's more of an emotional reaction; people say I wasn't included and maybe that's the question we ought to address. I don't think it's a less statistical problem, more an emotional problem to answer for the people who get uptight about it.

French: I just think the trend now is to sample the population as a whole, however, they fall. Or otherwise sample the plan that is representative of the population.

Hopkins: Just a question. Is there any real movement to develop like norms on retarded populations on like intelligence tests. I mean it seems...

Mann: Well not on intelligence tests but on achievement tests there was a call for awhile which suddenly quieted down. I think it wasn't a matter of norming but rather the feeling that usual achievement tests don't measure for the retarded child. There are now some new individual achievement tests. The Peabody Individual Achievement Test for one up. It certainly overcomes issues of distractibility to some degree, pencil and paper format and so on, but does it give us the information we want - it doesn't seem so.

French: But even on that test they didn't involve mentally retarded kids in the standardization. They were standardized on "the main stream of American Education", which was between

80 and 120 IQ, or the Peabody Vocabulary quotient.

Hopkins: The reason I raised the question, illogical as it seems to me was that in our institution the first few years I was there they used separate sex norms in GRE admissions. The 50th percentile in the quantitative test for males is the 75th percentile in females yet they were requiring the 50th percentile for any sex this seems taking it to an extreme case like to renorming an intelligence test for retardates. I mean I'm glad to hear that no one else is that illogical.

Mann: But there are parents who want to have the smartest retarded kid in the class, really I have encountered them. There is an important emotional factor. I think it is a very critical point that Larney pointed out, and, of course, for special groups there is the question as to the aggravations of the question. Rosa has made this point. Terms, same and different, and like and so on, certainly confused the results of many conservation studies when we found rather than conservations being tested children couldn't understand these terms, at least as the investigator intended. The same holds I guess for innercity groups on tests. On the Binet even identifying cows as "what gives us milk" is getting increasingly hard over the years for suburban populations. That is another issue of relevant content, Jay.

Millman: I'd like to meet the challenge of bringing down the level of our esoteric conversation and talk just briefly on advice on how to pick tests. I would probably answer questions as to how to pick tests in this way. I'd say first consider using tests instead of a test. That is to consider not just putting all you're eggs in one measurement basket but to consider perhaps different approaches to getting at the same phenomena, different methods. I think you'll have a richer kind of conclusions that you might be able to reach if you use multiple instruments. So that would be my first concrete advicd: to consider using different instruments. Second, I would say look at the reliability of the instruments. Especially look at the reliability of subtests. If you're planning to use and make statements about subtests. The reliability is a necessary but not sufficient condition it seems for the successful employment of the test. Recall that the reliability need not be as high for group and research uses as it is for individual prediction and screening, assessment, diagnosis. In assessing reliability pay attention to the group on which it was computed. If your group is markedly different it may be that the reliability may not hold up for your group. Now continue I would say the next thing as far as picking a test you have to then look at your purpose. I've identified three purposes that you might want to use your test for. My advice is a little different for each purpose. One purpose is for prediction and under prediction I would include screening. A second purpose would be for=testing theory, for building theory, theoretical format. The third purpose would be for description of people for the purpose of looking at the effects of intervention programs something like that. For the prediction purpose or screening kind of purpose then it seems to me that you want evidence on the extent to which your test has been successful as a predictor for the kind of thing you're interested in. You want what we usually call predictive validity. In the manual of the test they have to come through it seems to me with evidence that being able to reproduce diamonds or something is somehow related to whether they're going to become good readers or not, if that's the purpose you are usuing it for. You should look for that kind of evidence before employing that kind of test with that kind of purpose unless you're doing it on a purely experimental speculative kind of thing. Again you need to watch out what groups was such evidence based on if the predictive validity evidence is based on a need markedly different from yours the predictive validity may not hold up. If you're purpose is theory development then you're interested in what is usually called construct, and there you want information about whether the test seems to be relating to things that it should be. Also useful is kind of negative information is it not relating to things that it should have a low relationship with and this approach is suggested by Campbell-Riske Disc and their landmark article on convergent and divergent thinking. Convergent and discriminate validity is, of course, relevant for this kind of thing, like his arithmetic ability, his ability to make certain kinds of descriminations, things of this nature. Then it seems to me that you want to look hard at the items of the test. If you're building your own test obviously built on the work of others, use other test items from other tests if you can. If you can get a clear notion as to what is the universe of behavior in which you're interested identify items for the purpose of measuring if from other tests or develop your own. Let me conclude just with a statement about the relevance of norms for each of these purposes. I'd say that normative information, i,e, looking at how the test was standardized and how representative it is and so on, that norms, having good percentile ranks and all - that is not all relevant in prediction use, not relevant at all in theory use, and relevant in the description use only if you're interested in making comparisons with some other group. I would cememt that by doing that perhaps not too often are we interested in that. So therefore if the test does not have 9000 people taking it to establish norms, that may not be particularly important for your purpose. Vice versa, just because a test has expensive normative data doesn't therefore mean that it should be looked favorably upon. Large norm samples may be irrelevant as per its value for you.

Mann: Any comments on this or issues or corections or responses, Joe?

French: I'd like to ask a question about how one uses multiple criteria in developing their regression formula! Most often regression formula base several preditors on one criterion and how do you weigh the various tests that might be used as criteria to determine a prediction formula?

Millman: This is not clear cut. One cannot give a dogmatic answer here. People will give different answers depending on their persuasions. My answer is, if you must make a decision about a person, I mean a go, no go kind of thing, you have to have a single prediction equation. Will I or will I not put him here to have or there. Then it seems to me you have to make a weighting of some sort of your several criteria and I think this weighting should be done as much as possible on the basis of a value judgment, a subjective judgment as to the importance of these criteria that you're trying to maximize. This as apposed to something that would come out of canonical variant analysis. On the other hand if you're not forced to make a single decision about a purpose but you want to make separate predictions or if it's OK for you to make separate predictions about each of these criteria then I would treat them as separate criteria and predict them individually. Other people would answer differently.

Feldhusen: Well, I'm satisfied. That's a good answer there I wouldn't choose to discuss.

Hopkins: I like what Jay said except I think I would disagree with him a little bit about his comment about usefulness of norms as too general. I think an intelligence test without norms is of very little value.

**Millman:** Well, that's right for classifying, categorizing people. And to give different labels and so on obviously you must have a distribution and norms are relevant there. I guess I'd have to agree with Ken that there are more uses for norms that I originally gave credit.

**Hopkins:** I thought otherwise your statement was very relevant and very useful.

**Feldhusen:** One further note to repeat a point I made earlier. I'm still of the opinion that there's excessive acceptance of paper and pencil measurement and I previously suggested looking to other, perhaps, mechanical ways. Now the final note I make is that I don't take the book by Webb, Campbell, Swartz and all those people as a joke although their book, The Unobtrusive Measures in Education is a very humorous piece of reading but we have taken very seriously the idea of seeking criterion measures as well as predictors from that one category that Webb, Campbell, Swartz and Seacrest call the archival approach to finding unobtrusive measures. In light of your comments that schools are getting weary of testing now that's one good reason to seek alternatives to pencil and paper tests. I think the much more important issue is that among quite a few of the unobtrusive measures I have seen are things that are quite meaningful and emerging right out of the life or important aspects of the life of the child, or whomever it is that your assessing. I think potentially something that has great meaning, great significance, is likely to be found particularly among archival type information or other types that they've described. I close by urging that the whole thing be broadened here to include all types of measurement and a quest for unobtrusive measures as well as pencil and paper measurement.

**Bracht:** In regard to the comments John made I think there has been some recent interest in finding alternative measures to paper and pencil measures. There's one reset chapter in the education measurement book on performance measures, one author has an indication of this. There are also two publications from Education Testing Service on the use of performance measures on occupational education. Shinberg and someone else have been writing in this area which I think is a help in indication of some measurement procedures.

**Millman:** Performance measures are not without problems however, and they may not be a panacea attractive as they certainly are to many of us. I might make a final point about prediction studies. That is, that when we find that certain variables can predict certain later events, I think we have to be careful not to shrug our shoulders and say, "Well, that's the way it is; people of a low socioeconomic status are more likely to end up with this kind of criminal behavior." If it's true I don't know. There's something that almost bothers me about that and although we finally can predict certain things maybe we should be looking at what happens when we can change things around too. Our goal maybe shouldn't be just prediction. Maybe the best prediction variables to use shouldn't be those variables that we can not change. Somebody help me out please on this point!

**Feldhusen:** I want to speak about this point now that Jay has brought it up. The first problem is that, yes, over a long period of years people became very pessimistic about academic prediction research because it became a "so-what" kind of a proposition. So you got a multiple point 65 and you set it and that was the end of the story, or you cross validated it. People seemed not to go beyond. We made a quest of the literature, trying to find evidence that people went beyond this particularly in intervention ways and we were so hard pressed to find examples. There were very, very few instances where anyone did attempt to make use of prediction in other than the selection route, in colleges and universities for example. So we have now in progress a study where we are trying to beat our predictions or as it were to put our predictions or as it were to put our prediction system out of business. This is being carried out in the field of nursing education, where we're trying to make predictions. I would also suggest people could not understand academic prediction studies in the past because they seem not to understand and still don't seem to understand that you are talking about making predictions for person to person. I find this concept a very difficult one for many. They somehow don't believe, "You don't mean that you are going to predict for him and him and him." Well yes we do if our efficiency is high enough. So this is the very thing we are now about in a study: trying to make such specific predictions and then particularly to select out people for whom the prediction is adverse academically and to intervene in a way to prevent what is otherwise predicted.

## REFERENCES

Anderson, R.C. How to construct achievement tests to assess comprehension. *Review of Educational Research*, 1972, 42, 145-170.

Bloom, B.S. Testing cognitive ability and achievement. In N.L. Gage (Ed.) *Handbook of research on testing*. Chicago: Rand McNally, 1963. Pp. 379-397.

Burke, K.S., Nilges, T.C., & Henry, G.B. Middle ear impedance measurements. *Journal of Speech and Hearing Research*, 1970, 13, 317-325.

Catlin, J. In defense of sophisticated-guessing theory. *Psychological Review*, 1973, 80, 412-416.

Chapman, L.J., & Chapman, J.P. Problems in the measurement of cognitive deficit. *Psychological Bulletin*, 1973, 79, 380-385.

Diamond, J., & Evans, W. The correction for guessing. *Review of Educational Research*, 1973, 181-191.

Echternacht, G.J. The use of confidence testing in objective tests. *Review of Educational Research*, 1972, 42, 217-236.

Elkehammar, B. A look at "Another look at the reanalysis of the multidimensional scaling of Rorschach percepts. *Perceptual and Motor Skills*, 1973, 37, 266.

Elwood, D. Validity of an automated measure of intelligence in borderline retarded subjects. *American Journal of Mental Deficiency*, 1972, 77, 90-94.

Fiske, D.W. Can a personality construct be validated empirically? *Psychological Bulletin* , 1973, 80, 89-92.

Fletcher, S.G. Time-by-count measurement of diadochokinetic syllable rate. *Journal of Speech and Hearing Research*, 1972, 15, 763-770.

Gengel, R.W. Practice effects in frequency discrimination by hearing impaired children. *Journal of Speech and Hearing Research*, 1969, 12, 847-855.

## References

Holland, J.G., & Doran, J. Instrumentation of research in teaching. In R.M.W. Travers (Ed.) *Second handbook of research on teaching*. Chicago: Rand McNally, 1973. Pp. 286-317.

Irwin, R.B. Consistency of judgments of articulatory productions. *Journal of Speech and Hearing Research*, 1970, 13, 548-555.

Jackson, J.H. Measurement of ethical values. *Perceptual* and Motor Skills, 1973, 36, 1075-1088.

James, L.R. Criterion models and construct validity for criteria. *Psychological Bulletin*, 1973, 80, 75-83.

Kirkland, M.C. The effects of tests on students and schools. *Review of Educational Research*, 1971, 41, 303-350.

Linn, R.L. Fair test use in selection. *Review of Educational Research*, 1973, 43, 139-161.

Lord, F.M. Testing of two measuring procedures measure the same dimension. *Psychological Bulletin*, 1973, 79, 71-72.

Margolis, R.H., & Millin, J.P. An item-difficulty based speech discrimination test. *Journal of Speech and Hearing Research*, 1971, 14, 865-873.

Marshall, N.R., & Goldstein, S.G. The maintenance of diagnostic information imparted by three methods: One year later. *Journal of Speech and Hearing Research*, 1970, 13, 447-448.

Mermelstein, P., & Maclean, D.J. Computer-assisted measurement system for X-ray films of the vocal tract. *Journal of Speech and Hearing Research*, 1973, 16, 282-285.

Nelson, D.A., & Chaiklin, J.B. Writedown versus talkback scoring and scoring bias in speech discrimination testing. *Journal of Speech and Hearing Research*, 1970, 13, 645-654.

Overton, G.W., & Scott, K.G. Automated and manual intelligence testing: Data on parallel forms of the Peabody Picture Vocabulary Test. *American Journal of Mental Deficiency*, 1972, 76, 639-643.

## References

Owens, E., Benedict, M., & Schubert, E.D. Further investigation of vowel items in multiple-choice speech discrimination testing. Journal of Speech and Hearing Research, 1971, 14, 841-847.

Porch, B.E. Multidimensional scoring in aphasia testing. Journal of Speech and Hearing Research, 1971, 14, 776-792.

Proger, B.B., Cross, L.H., & Burger, R.M. Construct validation of standardized tests in special education: A framework of reference and application to ITPA research (1967-1971). In L.Mann & D.A. Sabation (Eds.) The first review of special education. Philadelphia: JSE Press, 1973. Vol. 1, Pp. 165-201.

Proger, B.B., & Mann, L. An historical review of theoretical and experimental literature on the teaching values of informal (non-standardized), teacher-made achievement tests: 1913-1968. Norristown, Pa.: Montgomery County Intermediate Unit No. 23, 1973. Available as ERIC document ED ******.

Remmers, H.H. Rating methods in research on teaching. In N.L. Gage (Ed.) Handbook of research on teaching. Chicago: Rand McNally, 1963. Pp. 329-378.

Sattler, J.M. Intelligence testing of ethnic minority-group and culturally disadvantaged children. In L. Mann & D.A. Sabatino (Eds) The first review of special education. Philadelphia: JSE Press, 1973. Vol. 2. Pp. 161-201.

Schoneman, P.H., Cafferty, T., & Rotton, J. A note on additive functional measurement. Psychological Review, 1973, 80, 85-87.

Shelton, R.L., Arndt, W.B., Wright, V., & Elbert, M. Item context and /s/ phone articulation test results. Journal of Speech and Hearing Research, 1972, 15, 852-860.

Shriberg, L.D. Articulation judgments: Some perceptual considerations. Journal of Speech and Hearing Research, 1972, 15, 876-882.

Stern, G.G. Measuring noncognitive variables in research on teaching. In N.L. Gage (Ed.) Handbook of research on teaching. Chicago: Rand McNally, 1963, Pp. 398-447.

## References

Stone, L.A., Gunn, S.R., & Lindem, A.C. Another look at reanalysis of the multidimensional scaling of Rorschach percepts. *Perceptual and Motor Skills*, 1973, 37, 257-258.

Turner, C. Effects of race of tester and need for approval on children's learning. *Journal of Educational Psychology*, 1971, 62, 240-244.

Wang, M.W., & Stanley, J.C. Differential weighting: A review of methods and empirical studies. *Review of Educational Research*, 1970, 40, 663-705.

Witter, H.L., & Goldstein, D.P. Quality judgments of hearing aid transduced speech. *Journal of Speech and Hearing Research*, 1971, 14, 312-322.

Woods, E.M. Recent applications of computer technology to school testing programs. *Review of Educational* Research, 1970, 40, 525-539.

Woo-Sam, J., & Zimmerman, I.L. Note on applicability of the Kaufman formula for abbreviating the WPPSI. *Perceptual and Motor Skills*, 1973, 36, 1121-1122.

Young, M.A. Anchoring and sequence effects for the category scaling of stuttering severity. *Journal of Speech and Hearing Research*, 1970, 13, 360-368.

## CHAPTER 11

## INNOVATIONS IN TESTING:

## CRITERION-REFERENCED MEASUREMENT

Overview: While standardized testing apparently will always hold utility for diagnosing relative degree of handicapping conditions, many question the appropriateness of norms-based measures for assessing day-to-day progress for exceptional children. A possible alternative to standardized measurement is sought: criterion-referenced measurement (CRM). In contrast to standardized, or norm-referenced measurement (NRM), whereby a child is compared to performance of a similar group of children, CRM emphasizes the child as an individual and looks only at his absolute level of performance, rather than his relative standing. CRM is also known under the headings of mastery testing and domain-referenced testing. Because in CRM one has no norms by which to judge a child's performance, he must determine what a pass/fail cutoff point will be; herein lies one of the major sources of criticism of CRM: the arbitrariness of the cutoff point selection. Other unresolved questions of CRM concern assessing the reliability and validity of the measures. Regardless of the controversy, however, CRM may be ideally suited to diagnostic/prescriptive teaching and other modes of individualized instruction as regards the teaching of basic skills.

know, ones for most needs that we might have. The problem is to go through them and dig them out. The question of objectives, of course, has been argued. Edgar Eisen, and others, argued that we don't need this kind of approach directly. Certainly not for social studies and art, but more so for math and science. The appeal of objectives lies in the implied objectivity of the material. Some of the other people argue about changeable objectives in social studies, but I don't know that that is a big argument.

Mann: Ken, do you want to add anything to the objectives?

Hopkins: On that particular point?

Mann: Or generally.

Hopkins: Well, I had some personal problems about getting on the criterion-referenced bandwagon. First of all, I think Glenn has made an important point—whether it is training or education—whether there is the element of transferability in the criterion measure that is important. I think if you are training someone to tie, or to keypunch, or to drive a car, I think you can establish meaningful external criteria. The problem I had was how do you establish that criterion—that magic criterion—on reading comprehension? I can easily get almost any group to achieve 90% criterion. Just write very simple items. "What was Billy's sister's name?" That is reading comprehension...and the arbitrariness of the difficulty. Now, the criterion-referenced advocates groom their measurement to their instruction, which I think works against the old transferability issue, in a broader sense. Another point is, that I don't think the conceptionalization is anything new—I think it is a new label. I think that Davis pointed out, in theory, it is not that different from the old percentage grading. Teachers fifty years ago gave tests, and those that got 90% and above, got "A's." I mean, that is the criterion. The real issue is, is that criterion a meaningful point? I think you can make a strong argument if your criterion is 100% mastery. But, that is unrealistic unless you are dealing with a very limited area.

Stanley: Ken, there is a problem with 100%, too.

Mann: I'd say 100% could be dealt with, in certain kinds of skills.

Stanley: Yes, there is a problem with that; anybody in experimental psychology knows what that is—can you do it two days from now? Can you do it five times in a row? Or, did you happen to reach it just one time. So there is, even with 100%, somewhat of a definitional problem there, although it is less than, say 90%.

Bracht: The problem that exists here is that people are using the term, criterion-referenced measurement, often have different ideas of what that means. I am thinking about the chapter in Educational Measurement by Glaser and Nitko. And of course Glaser is, I think, frequently identified with the criterion-reference of measurement. They refer to a measurement of a very specific skill; for example, adding one-digit numbers. Here, the content domain is very identifiable, and also the measurement procedures are very clearly identified. We can use these as a vertical format, or the horizontal format. And, we could also do this, say with two one-digit numbers, or three one-digit numbers, or something of this sort. So that the content domain is clearly identified. If we give a person a task, and the task consists of adding ten of these problems, then the score is—well, I would predict, although I haven't seen any results—that you would problably get either extreme low scores or extreme high scores for most of the examinees. Then, you can readily interpret the performance. As we define the content domain more broadly, say the content that mainly comes with arithmetic ability or achievement, in addition. Let's take that...achievement in addition. I think it becomes more difficult to have a criterion-referenced measurement, or if we take reading comprehension. I don't think we can get a criterion-referenced measurement in that area.

Mann: You can, if you live with curricular-scheduled objectives—if you stick specifically to the content of a particular content. As Ken says, it is problematic that this will generalize to the generic global construct of comprehension.

Bracht: One of the misinterpretations has been: If I set a criterion of performance, you have to get 70% of the items correctly; you have to get 90% of the items correctly; then we have criterion-referenced measurement. That is not the sense in which Glaser and other advocates of—in that camp—are talking about criterion-referenced measurement. That is just setting a standard of performance—a criterion for performance. That has been one of the confusions, I think. If it is a very specific skill so that the score obtained is directly interpretable, I think—you know, like a typing test—this is a criterion-reference of measurement. I think there is a use for that type of measurement, but I don't think we can now say that criterion-referenced measurement replaces other types of measurement. The diagnostic testing that Julian refers to, in my mind, is criterion-reference of measurement. We measure a very specific skill; we are going to see if the student has mastered that skill.

Mann: Could you reiterate Glaser's definition for our audience?

Bracht: Well, I can't repeat his definition, but criterion-referenced measurement is used in the situation where the performance is directly interpretable. So, the score that we obtain for a student's behavior is directly interpretable.

Mann: This is relative to competence in that behavior. Is that correct?

Bracht: Yes, we can say he is competent, or that he is not competent, in that skill. For example, a score of nine on a ten-problem test adding one-digit numbers, would indicate that the student has mastered, or reached a competent level of performance; while a score of maybe one or two, would indicate that he has not mastered that skill. But, if we give a test in general reading comprehension, or a test on general addition ability, a score of twenty is not that interpretable because people with scores of twenty can obtain them in different ways. One person may obtain it by doing very well on certain types of problems; others may obtain that score (twenty) by performing other types of problems—they don't all have the same levels of skills mastered.

Hopkins: I would like to get Glenn's reaction to the thought that it seems to me, that criterion-referenced approach becomes a much weaker model the minute you move beyond the knowledge level, so to speak, in Bloom's taxonomic structure of different levels of thinking.

Stanley: Very important.

Bracht: I agree. I think we can use CRM for skills and basic knowledge.

Mann: We believe it is valuable in special education, just for that purpose. It is really a quite modest approach. Jay?

Millman: Yes, I'd like to make a couple of speeches. They will be shorter, because of the comments of Ken and Glenn because I agree with them—much of what they say. Some people have used the term, domain-reference, for criterion-reference, to help avoid the confusion. I think the critical thing is, can you specify an item generation procedure? If you can specify an item generation procedure, then you have a domain-reference of test, and if you want to put on some sort of passing score, that, I guess, you can call a criterion-reference of test. It is certainly true that it is harder to generate items and develop generation procedures for domains concerned with complex skills. Although I think you are just a little too pessimistic, as to its limitation. Certainly discriminations—discrimination learning—you can, if that is your interest, certainly see if people can learn certain kinds of discriminations which go higher in Bloom's Taxonomy. My second speech goes back to this objectives thing. I took a leave of absence from Cornell this summer, and was on the payroll of the Instructional Objectives Exchange—so, there is a conflict of interest here, you know. I was interested in

Rosa's comment, that she noticed that the teacher in her teaching, is really operating on a very limited number of alternatives. And it is my impression, that she also has a limited number of objectives, and a limited number of ways of measuring, if she has been successful in those objectives. I think one of the hopes for the Instructional Objectives Exchange was to present to teachers: a wider array of objectives for her to consider, than the ones that she may have thought about; to open up alternatives of objectives for her to work on; and, to open up different ways of measuring whether they have been successful. I think even our professional test constructors know that the novices—and I will put most teachers in that category—I think have tried to test whether they have gotten something across...in very narrow and restricted ways. Unfortunately, I think the initial collection of objects that I have seen developed have been of very poor quality. However, in the revisions that are going on now, they will be better organized and, I think, show—hopefully—a greater range of types of items that could be used to measure meaningful teaching goals. These items—ideas about items—could then help the teacher to make her own little diagnostic tests, or domain-referenced/criterion-referenced testing instruments for her own evaluations. My remarks, however, do not deal with Joe's concern of which of these objectives should she teach—which is another whole bag—and the purpose of these Exchanges is, I gather, not prescriptive, but merely to give choices for school administrators to select from.

Mann: Jay, when you say that the teacher has too narrow a set of objectives, are you inferring that they are too vague, or that they are just focusing on just minor subdomains of behavior.

Millman: Well, for example, to use an illustration that I used before; she may think, well...I want a kid to learn how to add simple two-digit numbers—you know, that is my objective. Really, when you get right down to it, all she does is give him practice, when they are neatly lined up; they are all aligned; when they are in the vertical position, and they don't consider all the kinds of sub-objectives that might be in the different ways you could test the kid. It just doesn't occur to her that there are some other things to do—that she might teach them—as per set training in, and the new math, and she might not really think about getting the difference between equivalent and equal sets across, given a bunch of objectives to choose from. She might say, "Oh yeah, I wasn't going to teach anything about that, but yeah, here is this objective. Yeah, I think that is something I want the kids to know about, and I hadn't thought about it." I might help in those ways.

Hopkins: I just think what Jay says is really...I really agree with

Mann: Perhaps we can now discuss some issues in diagnostic and criterion-reference testing.

French: We have one thing that bothers me, and that is we have so few useful diagnostic tests.

Stanley: We actually have a lot, but we don't use them. We don't even know about them. They exist in, say arithmetic—commerical—arithmetic: Southwestern Publishing Company. I used to teach that course many years ago and there were lots of good diagnostic tests available...plenty of them.

French: The thing we special educators want to know is how do we know that the published tests cover all the objectives that ought to be covered in the program.

Stanley: Well, the diagnostic ones do a thousand times better than the teacher will do. Many are put together in some very systematic way along the old lines of Stow and Curtis, and so forth. We also know that the teacher is very uneffective oftentimes in devising his own diagnostic tests. If you want to know whether a kid can add three-digit numbers in sets of five, you can get tests of three-digit numbers in sets of five. The workbooks are full of them.

French: But what about—aren't they designed more for a particular test or program and fail to deal with those items that Glenn was talking about that students ought to pick up on their own while they are in a classroom, that the teacher doesn't teach more directly? Now, maybe in a class for the mentally retarded or class for the physically handicapped, the teacher has to teach for each of those items directly. But, the thing that bothers me is that I don't see the teachers' knowing the objectives that ought to be there. I don't know the objectives that ought to be there. I don't know that the instructional objectives exchanges that I have seen from Westinghouse or UCLA and elsewhere are right. You know, they are different, and you see a bunch of objectives. Are these all of the objectives, or are these the more important objectives? We know that they are not in an hierarchal order. How do we go about determining the kinds of objectives that ought to be known by the teacher, so that she can guide instructions in those directions? How does she arrange them?

Stanley: Well, you have a number of ways. The test, in effect, oftentimes states the objectives. Now, there are different kinds of tests. I was talking about diagnostic tests versus aptitude tests, versus survey tests. The general science knowledge test pretty obviously doesn't you know, limit itself to whatever teacher happened to teach, and a fourth type would be the teacher-made test, made up of different things. We have lots of instruments available—thousands of them—and I think that among the ones already available there are, you it. I think he reinforces the point about even when we can specify the domain—like two-digit numbers—it is still kind of an artifactual...do we define it in terms of whether those are even lined up correctly; or, does the kid have to write them down; or, whether his alignment could enter into it; or, whether he does it orally, at the blackboard, at home, or altogether.

Millman: As a postscript, the objective exchange I was looking for is for the affective areas, where the teacher, for example says, "I like your self-concept objectives." By the way, I really hope there is a self-concept. Well, the exchange is now describing different suggestions for ways that you can measure whether a kid's self-concept has improved. The first we've had; I hope the second batch will be better.

Mann: The second ones will be bad, too. I can almost guarantee you.

Millman: Yes, but they are probably better than what the teacher could come up with on her own.

Stanley: It seems to me that, ideally, diagnosis can never be done with just any kind of test—no matter how good—alone. You have got to have interrogation—you have got to go back to Bloom and Boda —asking the youngsters what they are doing; trying to determine what procedures they used. We have an illustration with one of our bright sixth-graders that, I think, points it out. In the SCAT college level, there occur exactly two items that, on the mathematical operations, involve dividing one proper common fraction by another, you know, 2/3 divided by 1/2 type problem. Well, these happen to be two problems. Now, had our youngster got the two of them right, we would assume that he knew how to do it—but, he got one right and one wrong. So we talked with him about it, and he said...well, he had figured one of them out in a very clever way. He didn't know the rule—he didn't know the invert and multiply rule. But the first one was accessible, but unfortunately, the second one was too difficult to use his partitioning method on, so he missed it. So we learned that he had to be raught the rule and the reason for the rule, which was a fairly quick proposition. Now, what I am saying is, that diagnosis had to be supplemented here by the general question of, "What did you find hard in the test today?" The first time he took the test he said, "well, you know, I found dividing one fraction by another the hardest. Why? Well, I don't know the rule." And so forth. Another illustration we had was a girl who was able to add two- or three-digit numbers—a list of them, five or six—two-digit numbers —but did it very slowly. She happened to be a Black senior; a very bright girl from—actually a part of Virginia—in a segregated elementary school. And we discovered, finally, when we began to look at her work, was what she did was to add two of them at a time. She didn't know a whole series, like six; she added the first plus the second and that answer plus the third, and although she could do it, she did it very slowly. Now, again, if we just gave her plenty of time—gave her a number—she could add most of them correctly. It was obvious that we needed to investigate. So, diagnosis means more than getting a criterion-referenced type score. It means, ideally, looking at the middle operations. The third, and simple example, is, a very clever kid can figure out—say arithmetical reasoning problems—even though he is not very good at fundamentals, by using something level four of Bloom's Taxonomy Analysis; whereas, a person who is older and more trained and taught, is probably using level three, at most, to do the same problem. And, we don't learn about that unless we have some kind of Bloom-Boda interrogation methods to talk with the kid about what he is doing...How did he solve this problem?...Did he use this level, this level, or this level?

Bracht: I am reminded of another comment here, regarding the use of positive taxonomy. Sometimes, we see a study in which a set of items are classified as being knowledge items, comprehension items, application items, analysis items, etc. And, I think there is value in doing this sort of thing and I don't want to discourage its use, but, I think we ought to be aware that we really can't take an item and say that this is an application item. Or, we can't take an item and say, this is an analysis item, because for one student, it may be; for another student, it may not. Let us take an item which requires a student to synthesize various pieces of information to construct something. Well, for one student, that task may be purely a synthesis-task, but another student may have read something else previously, in which this very thing was done, and so, when that student does the problem he is not synthesizing, he is just recalling what someone else has done. So, he doesn't have to synthesize at all. So, I don't think that we can classify those items according to that taxonomy of the complete extent. I think we have to look at the process that the examinee goes through when he does that task. So, again, I think for the most part, we are pretty successful classifying items, but there are some problems.

Stanley: A striking example would be the type item: John is twice as old as Mary was ten years ago. If a person knows a little algebra, he can zip those off. If he doesn't know any algebra, he can still solve them, but he is using a high-level process to solve them. He has to figure them out a different way.

Millman: We need to look at the previous educational history of the student in order to assess what the item is measuring, and it also delves into the problem of creativity, especially when you think of a particular response as being original, or creative, or something. It is hard to just look at the response and say, "Ah, there is an original one, or creative one," without knowing the history of that child.

Hagin: I found, in my practice with children who have learning problems, that it is as much important, now, to get a child's educational history, as to get the developmental history from the mother because the approaches to teaching subject matter, for instance as mathematics, will determine what a child can do. Sometimes, they have not handled certain operations, and where other operations have been much more emphasized, since the current standardized tests may sample items that don't reveal this, it is very important to me to see in actual examples of what they have been taught in order to evaluate, whether they had the opportunity to learn it and they failed to learn it, or whether they were ever exposed to that particular thing. This business of asking how a child solves a problem—it is very interesting to see whether they say move to another modality to solve a visual problem, or whether they use verbal mediators—unless one asks, and unfortunately, one doesn't find this out very often.

Millman: I would like to make a summarizing statement about this whole business of domain-referenced or criterion-referenced measurement. In application for, say special education, especially where you are trying to teach some cognitive skills to youngsters, I think educators are now on the bandwagon by being very concerned about the sequencing and prerequisites—what the student knows, and so on—to take into consideration where he is, in order to determine what to do next. I would see a concept emphasis, in special education, on tests that don't try to compare individuals with other individuals—the old normative tests—but increasingly, an emphasis on how does the kid do on items that are, in a sense, representative of a particular level or kind of skill—CRM criterion-referenced testing awareness to a very definite kind of role in special education. It seems right in line, you know—in tune with the philosophy—we try to look where the kid is, and try to have the next instruction be logical, based on where he was. It just seems to fit in very nicely with that philosophy.

Cross: I quite agree with everything that has been said. I think, being a student of Fred David, I can't help but agree with what has been said. However, it bothers me that people seem to think that criterion-referenced measurement is such a good thing. I have a bias, and maybe this has been expressed—I think it has—it is just great, in terms of achievement, or where you...criteria is definite and it is definable, but when you move over to the measurement of constructs—where you don't have a unitary, single criteria to measure—I think its applicability goes down.

Mann: I think Glenn said that very clearly over there, and we all agreed.

Cross: But, in terms of when do you use it?...I do think there are some technical problems that Livingston has addressed himself to. I think there is some problem when he replaced the group mean by the criterion level, and he talks about reliability—about this criterion level—as opposed to group norm. And, everything goes along beautifully until you stop and think that the optimal goal, in norm-referenced testing, is to maximize dispersion about the mean, in many instances, and yet criterion-referenced measurement—that isn't the case. You want to bring everybody up to criterion, and you want to minimize that variance. What happens to reliability when you are trying to minimize that kind of basic ability, you know, as Livingston has defined it—what happens to, when you talk in terms of validation...validity of items, and so forth, when you know you are getting limited variance. You can't talk about predictive validity of items; you can't talk about the construct validation of items, or tests—your variability is optimally decreased.

Hopkins: I think the notion, or reliability, needs to be approached in a different way for CRM. It is more...would you make the same decision on a parallel set of items? It is more of a go, or no-go, on two occasions. How consistently did you repeat the same information? Is it above 90% both times?

Millman: Or, if you don't take the criterion that seriously; that is, the cutoffs for that test, seriously, then the crucial question in domain-referenced testing concerns the standard error of measurement, which is: how did he do on another parallel testing? The standard error of measurement is the same, I believe. This is one thing that Harris pointed out; it is quite correct. The standard error of measurement is the same whether you compute it, using the CRM reliability, in the... isn't this right, Julian?

Stanley: It is terribly important how far you are from that criterion point, with respect to the standard error of measurement, rather than the mean. That is the thing that I think that Harris is not emphasizing. Clearly, if the criterion is here —and you are here—then you are six standard errors of measurement away from the criterion score. Then, you aren't ever likely to reach that criterion score because of errors of measurement on repeated testing. I think, in that sense, the projection is not valid to that. The real thing in Livingston's sense of his argument is, algebraically, impractical. Everybody agrees on that; they have checked it; I have checked it. And, other people have checked it—Chester Harris has, for instance. The real thing hinges on how you define criterion-referenced measurement. If you accept Livingston's definition with the C sub x, then you can't reject his system; it is all there. It is, however, algebraically, and deductively, circular. You don't have to accept that definition, since there are many definitions of criterion-referenced measurement.

My own personal feeling, as a person who has worked with Livingston—somewhat checked out his work—in particular... although I didn't suggest it to him; it was his own idea... is that CRM is another semi-false Holy Grail—another fad. And, it is going to be pushed to death, like all fads are, and it will leave a small residue—a useful residue, later on —and that will be it. It does not solve any great problem of measurement, or anything else; it probably focuses better on clearer specification of objectives and attainment, but you are losing a lot of information, and many uses of it, by throwing away the metrics and taking a dichotomous stand.

Mann: I think the reason we moved to it, in special education, in certain corners—Bart, Larny, and I—is that curricula in special education has not been a success, and we have proved, if nothing else, CRM establishes clearly defined instructional objectives in limited areas—that the kids can reach these and achieve them. And if it did nothing more than that, CRM itself, is not terribly critical to our system, but it does focus on the instructional objectives. Joseph, if you will?

Stanley: Well, in that sense, I think it is fine, but I think it has some kind of psychometric panacea to it—overrated.

Cross: I think the problem is the abuse, or misunderstanding, of it. I think it is not going to replace norm-referenced testing and where norm-referenced testing has been used appropriately, it is useful in individually prescribing instruction, and where we have a very thorough task analysis, with a very specific task, we want to know if a student is competent in that task. We haven't used norm-referenced tests for very specific purposes like that. So it is really not replacing anything that we have done before, but is bringing something new which I think is useful.

Stanley: Old fashioned pessimism on your part...still!

Millman: Julian has the advantage of a little longer view of things than I have perhaps, and can put these things in a better perspective. But it does seem to me that the aspect of criterion-referenced measurement where it is a go, no-go kind of thing, may not have the utility some people hope, but there are other things that I think will linger longer. One thing, that I think that it will leave as a legacy, will be the hope of people to try to specify more clearly the domain of skills that are being assessed, and the ways in which items are generated to sample from that domain. I think that will be something that will stay with us.

Stanley: CRM does tie together a lot of things...you could throw Guthman's facet theory into the pot with it. You can talk about factorial pure test. You can do a lot of things there that are usefully put together in one framework that haven't been done before. It is not a new concept. However, I know that when I was teaching chemistry as far back as 1938, I had to be eternally concerned with getting those lazy seniors that I had—boys, particularly—to learn to associate the names of the elements with the symbols of the elements. We couldn't make any progress in using formulas until sodium was automatically Na, and Na was automatically sodium. They could do it orally, visually, and write it, and talk it and everything. And so, in a sense, you had a criterion-referenced problem, even then, in that situation. But it was impossible, of course, to get mastery from everybody at the same time. We'd spend half of the year trying to get everybody up to twenty out of twenty symbols mastery.

French: Well, I would like to get back to the concern of the teacher in developing evaluation criteria. If we go to criterion-referenced measurement more than we have, and advocate it more than we have, it is going to create either tremendous increase in need for published materials, or a training need to help other individuals within a school system construct evaluation devices, so that we measure a particular concept. As we were talking about—with the arithmetic variable—not just two digits directly beneath each other, but the various ways that are important in determining whether the person really has that concept. So it creates, as I see it, an enormous evaluation problem that we haven't even recognized, I think, in many instances; most standardized instruments just won't meet the bill. So, it may be a whole new field for construction and development, in special education.

Cross: I think it would probably have to be a part of the curriculum packages—for example, the triple Ss curriculum for elementary schools. It has the lesson and it has the short test that is used to see if the student has mastered that skill, and that is what we have done for the Palo Alto curricular.

French: Because there are so few handicapped kids, it is going to be a long time before it is commercially feasible to enter into that kind of activity for handicapped kids.

Hopkins: I think I agree in what Julian way saying. In special education, however, I think particularly in the TMR classes—so many of the objectives are training uses, rather than educational uses: tying shoes, toilet training—these are things that I think fit very nicely into a criterion-referenced strategy.

Mann: ADL, they are called—Activities of Daily Living. Go on.

Hopkins: And so, it is fine in those kinds of situations, but I think what has happened is that some people think that this is to supplant norm-referenced testing, and all that. Well, I can't see that happening.

Stanley: The experimental psychologists...I keep coming back to the notion of learning to a certain criterion. For instance, can a youngster tie his shoes? Well, you are asking, in effect, can he or does he, in nearly any occasion when he is supposed to tie his shoes, tie them properly? Tie them so that will stay tied? Performance to a certain criterion—not can he once reach up to that level and get the shoes tied one time—but is this a customary thing that he does? CRM measurement, somehow, has to accommodate this.

Mann: I want to make a very fast go-around. I'd like to ask this question of you, Julian. We discussed the minority measurement issue, and how we will probably never resolve it, in terms of anyone being happy. What might happen when it comes to a CMR system for said purposes are not longer in any unfair comparisons. We say, okay, here are the criterion levels you agreed you needed in areas of achievement; everyone should reach their black-white checker. Would CRM be a possible way of overcoming the minority hostility to testing? The anti-testing climate is very rapidly catching in this country. Big cities have eliminated much of their testing programs. Now, would CRM be a way out of this for socially useful purposes? And, of course, CRM can show progress, even if it is slow.

Stanley: But the problem would still be there. If the youngster's learning potential is a good deal lower, there will just be a larger percentage of them who won't reach the criteria. You don't get away from that problem very easily, unless you put the criterion down to some absurdly low level.

Millman: I answer the question this way. Criterion-referenced tests of the type you suggest would not probably be considered racist or biased by minority groups. However, I think parents—black, white, yellow, or green—want to know, not only what can my kid do and to see his progress, but are also interested in how does he compare to other people. I don't think we are ever going to get away from the need also to sometime answer the comparison question.

Stanley: We implicitly do it, by saying in effect—even college admission—how does he compare to other blacks? You know, in effect, we oftentimes have two different standards for admission to college. One is the standard: is he at the 98th percentile of blacks, even though he is only at the 22nd percentile of whites? We don't admit that, or acknowledge it publicly, but it is implied in admissions procedure.

French: One of the problems of late that we are not considering, is the amount of time it takes to reach that criterion. Also, what are the implications of teaching for the test, as we were talking about earlier. The test is no longer as predictive, where there has been a great deal of instruction leading up to it. We also want to predict how well children will learn beyond the criterion point. Will they be able to pick up the skills in the industry, or will they be able to profit from the instruction in the college, or whatever?— not have they just reached this level of criterion—and how about the implication of transfer and the speed of later learning?

Hagin: Well, I'd like to take another tack, and that is, that we will have to learn very soon to be able to explain what we do in testing to parents, and to explain it clearly. We have reached the point where school records will be open, and it seems to me that criterion-referenced testing can be explained more easily, in terms of performance, than some of the norm-referenced tests. The parent can more easily understand the child's gains and lags.

## REFERENCES

Glaser, R., & Cooley, W. M. Instrumentation for teaching and instructional management. In R. M. W. Travers (Ed.) Second Handbook of research on teaching. Chicago: Rand McNally, 1973, pp. 832-857.

Millman, J. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205-216.

Proger, B. B., Mann, L., Burger, R. M., & Cross, L. H. Adapting criterion-referenced measurement to individualization of instruction for handicapped children: Some issues and a first attempt. Paper presented at annual convention of the American Educational Research Association, April, 1972, Chicago. Available as ERIC document ED-064-354.

Proger, B. B., & Mann, L. Criterion-referenced measurement: The world of gray versus black and white. Journal of Learning Disabilities, 1973, 6, 72-84.

## CHAPTER 12

## REPORTING PUPIL PROGRESS

Overview: An area often neglected in discussions of testing practices is the simple topic of reporting test results. Test reporting practices are becoming critical with the framework of the bigger issue of invasion of privacy, privileged information, and so on. Basically, there are two ways of reporting progress: (*a*) in terms of group norms, and (*b*) in absolute performance levels (Criterion-Referenced Measurement). Norms appear highly appropriate for traditional special education testing for screening, diagnostic evaluations, and individualized programing. On the other hand, absolute levels of performance are more suitable in research for comparing various treatments and in applied practice where one monitors on a highly specific and frequent basis the progress of a child. Finally, there are a host of issues (not all of which are addressed in this chapter) as to what types of reporting are appropriate to what types of people (school administrators, guidance counselors, parents, teachers, students, and the community at large).

Proger: I would like to suggest that we consider the problem of reporting progress. How would you suggest it is best to convey progress in an EMR child or a learning disabled kid, or even a trainable mentally retarded kid who is barely capable of tying shoes, and so forth? What would you suggest as a means of conveying to parent the problems of their child? I don't think standardized tests helps much where you have kids who are handicapped. I believe you probably need custom-made, very special down-to-earth level of functioning tests. CRM tests, traditional achievement tests that are just too global and guess, from the standpoint of the handicapped. I think you have to go much deeper into very precise type of behaviors somewhat alien to what Glenn was discussing earlier.

Millman: Well now, if you want to report to parents, I assume you want to report to a particular parent about his or her child, and in that case you want to give feedback to the individual parent about his child. I would think that that makes a lot of sense and talk in terms of what the kid now knows and can do that he didn't know or couldn't do before, and this would be done along single CRM lines. But, if you would rather want to talk about...to professionals in the field about what things can we expect of this type of person, then maybe a national assessment type thing would be more in line.

Mann: Would you say a few more words on that?

Millman: If your purpose is to gather information about what EMR's can do, then I think it makes sense to me to do it—provide the kind of data that national assessment does. I think that would be very useful. I think that is to indicate what percent of this type of population can be expected to do—these kinds of things.

Proger: Is that actually in a way of developing some indirect norms?

Millman: Yes. If that is what you are interested in—in getting norms. I think the definitional problem in special education needs serious consideration. If you want to say that 70% of EMR's who are of this age can do this kind of thing, then it seems to me—the purposes of communication among professionals in the field—you need to be very clear about what you mean by EMR's and give some assurance that you've got a representative sample of that particular population. I am not sure how helpful such data would be, but if you want that kind of data then you proceed along these lines. Now, what value is providing normative information about EMR's?...that might be an interesting question.

Mann: Well, a number of people have proposed that this is a legitimate question, really, to have a thing-norm for EMR's. Some people have suggested this and, indeed, have tried it. Joe?

French: Well, I don't think it is a bad idea—having the norms—but maybe perhaps what is more important is to determine those things that are necessary for job entry skills and establishing a criterion for various classifications that relate, not to specific jobs, but rather family clusters of jobs, and then maybe you just determine how many of those a person has, rather than what the norms might be for.

Proger: It would sort of be mastery again—perhaps absolute instead of relative mastery.

Hopkins: Well, I think that there are two different issues, so I am getting confused. One is strictly research—use of some measures—like you were talking about. We could have trainables that we do something to. We can only service ten and there are twenty who want the service, and we took half of them and then we came back and find out the percent who could tie their shoes in both groups, and all that. We could compare them and see if the treatment has efficacy.

Hagin: Well, there is another thought current, and that is the matter of reporting to parents is becoming increasingly one concerned with civil rights, and I think the time will soon come that anything that is on the school record or in the psychiatrist's record will be available, and must be interpreted, directly to parents. I think we are going to have to decide how to report clearly—objectively—anything we put down...so, it will have to be a matter of open record.

Mann: And provide a few inferential statements regarding presumed dynamics or relationships of parental pathology to outcomes in school, and so on, which will get us in trouble..and indeed has already.

Hagin: There is an APA committee, advisory, looking into a standards per testing, and it is just getting organized. I have a letter from someone who apparently is not in psychology, but who certainly knows what is going on, and there is one paragraph which says that his group—which he doesn't name—is a national group which will humanize testing, or destroy it, and then he wants to know what we, as school psychologists, are doing about the issue. Well, we are doing some things, but I am not sure we are going to be a party to this swing... but this is the way things are.

Mann: Jay?

Proger: Well, the ETS is using the situational tests—and where they examine actual behavior in the classroom—that don't have any implications, as far as being dumb or bright, or anything else, and then they can report to the parent, or whomever, what the child is doing in this particular situation. It is observational data and it is scaled, but I see it as being part of the Arithmetic Test Movement.

Hagin: I don't know this.

## REFERENCES

Coles, G. J., & Stone, L. A. Instructors' multidimensional perceptions of their students. Perceptual and Motor Skills, 1973, 36, 13-14.

Goldstein, K. M., & Tilker, H. A. Attitudes toward A - B - C - D - F and Honors-Pass-Fail grading systems. Journal of Educational Research, 1971, 65, 99-100.

Hammer, B. Grade expectations, differential teacher comments, and student performance. Journal of Educational Psychology, 1972, 63, 454-458.

Lindgren, H. C. Strong's psychologist scale and course grades in psychology. Perceptual and Motor Skills, 1973, 36, 58.

Means, R. S., & Menas, G. H. Achievement as a function of the presence of prior information concerning aptitude. Journal of Educational Psychology, 1971, 62, 185-187.

Schaefer, S., & Engel, R. R. Operant control of autonomic functions: Biofeedback bibliography. Perceptual and Motor Skills, 1973, 36, 863-875.

Snyder, C. R. Effects of comparison level feedback on classroom-related verbal learning performance. Journal of Educational Psychology, 1972, 63, 493-499.

Stone, L. A., Bassett, G. R., Brosseau, J. D., DeMers, J. L., & Stiening, J. A. Training staff's multidimensional perceptions of a class of Medex (physician's extension) trainees: A method of grading. Perceptual and Motor Skills, 1973, 36, 395-402.

Weissman, S., & Slonim, P. Effects of knowledge of results on dynamic stereo acuity in males and females. Perceptual and Motor Skills, 1973, 36, 964-966.

Wexley, K. N., & Thornton, C. L. Effect of verbal feedback of test results upon learning. Journal of Educational Research, 1972, 66, 119-121.

# CHAPTER 13

# RESEARCH UTILIZATION

Overview: One of the recurring problems in research and development efforts in education is how to have practical impact in the field. Part of this chapter discusses how to achieve greater utilizability of results. One suggestion is to form an agency that sifts through the findings of several studies and then forms a synthesis of practical implications from them. Another option is to form data banks to pool results in order to gain a composite interpretation of what the situation is in any given area. This chapter also discusses the ethics of "fudging" data or the application of Finagler's construct. Finally, the role of the educational R&D labs in the research information dissemination effort is described.

**Mann:** Are there any points about validity or any related topics?

**Feldhusen:** Apparently there is a great amount of dishonesty in some research domains—some caused, apparently, by naivety. An editor has just written to me about a manuscript we are working on, in which he suggests that we may be the first people to report research with the Gluck prediction scales who have not looked simultaneously and advantageously at their predictors and their criteria. I don't know if that is so; if it is, it is a devastating thing, because there is a fair amount of information in the literature already about the Gluck scales as predictors. The implication of the communication was that the looking at criteria and predictors simultaneously was, somehow or other, done in a way that enhanced the conclusions, or the kind of results, drawn from the study. So, apparently, there is the hazard of dishonesty. We spoke at lunch of Fred Davis' review and research of the prediction of intelligence and brain waves. His work suggests, if not downright dishonesty then, at least, fantastic naivety, on the part of brain wave researchers. "Fudge" is a good word, I guess..."fudge: is a way of being sweetly dishonest.

**Proger:** There is certainly a good deal of questionable research and lying around. What safeguards or what procedures can we use to even begin to hope to use? Also, what can protect us from the mass of inconsequential and unrelevant research that hits us?...that we, in fact, may contribute to ourselves. You look at small _N_ studies; they are cut and dry. They crank out a few results and then provide us a nice discussion, but as far as realistic classroom practice goes, little emerges. Maybe this is an inappropriate question. Maybe we should be saying it is basic research and that we are only slowly building up to the point where somebody can cry down from on high that it has some implications for practical education.

**Millman:** You argue, it seems to me, for a mechanism by which educational research can be made programmatic and by that I mean, a procedure by which all the little studies can be made to build on each other. Let me for a minute digress and tell you what—an exciting thing—that a colleague of mine is doing in the field of educational administration. He looked around and he saw the literally thousands of dissertations being written in educational administration, and the literally tens of thousands of manhours being spent on this research—most of which is not having any kind of impact, or getting any kind of definitive answers to questions. And so, he is attempting to marshall this tremendous manpower of dissertation-makers—this tremendous resource—to work programmatically on certain kinds of questions, and he is looking at the feasibility of making use of this effort. It seems to me that perhaps in special education, where _N_'s are small and where your problems are critical, that maybe there ought to be similar effort to marshall the available resources and to take problems and attempt to get students who are writing theses to work on assigned problems. If you can get that kind of cooperation on problems with available kinds of students from the variety of geographical areas, you might have some answers. It is done in medical fields and I don't know why, perhaps, it couldn't be applied with special education—an attempt to get more programmatic research tactics on some of these questions. You would increase the _N_'s on critical issues, replicate and generally answer some important questions.

French: Well, I think that is a very good idea. It would certainly be a vehicle for increasing comparison groups and the size of the _N_'s. I think that you've got a slight problem in getting thesis committees in various universities to accept projects proposed for dissertation by somebody else as an independent investigation for a particular thesis written in that department. But, if you got a funding for the thing, you could certainly at least employ graduate assistants to do this type of thing. It seems like a very feasible approach.

Bracht: I'm not bothered as much by the small sample size of a lot of these studies, as I am by some other factors. I think, so often, research is treated in such a simplistic fashion, you know—you give treatments A and B and you give one post-test. That, I think, isn't going to produce too much knowledge. I think we need to, in many cases, have multiple dependent measures. I think we need various systematic observations as to what occurs in treatments, and those who are administering treatments. We need systematic information about the participants in the experiment to stratify the population in more sophisticated ways. I think those things would contribute more to knowledge than just having larger sample sizes.

Hagin: It seems to me that also it would help things if the researcher would report in, or make available, as appensix or other material that can be mailed, procedures so that others can replicate his wish, or know what has not been done. I recently read papers submitted for publication which described intervention, in terms of counseling in psychotherapy in reading, and reports highly favorable results, but did not specify the kind of therapy that was presumed to have been done. It would be impossible to replicate this experiment and yet it provides very startling results, and where we get negative results we might be able to correct the study, or replicate it correctly, if we had enough data.

Feldhusen: Now, I am quite worried about re-doing parts of a study that turns out negatively. I have observed, in special education, that quite a few people operate from a position of discipleship under relative Messiah-type persons and their minds are made up, and they do research accordingly...beset by problems of the theories that are accepted as gospel, and on faith. This happens in other realms, too, where practitioners get to believe that somebody has finally brought the message from on high. But it certainly presents a problem, I think, in research, specifically, now, for somebody who is going to start tampering around with a study again, to dodge or to throw out cases that don't seem to come out right, or what have you. So, I'd be particularly worried. An area that suffered from this was ITA; ITA certainly was a realm that bred this same sort of condition—a faith—that here was something from almost on high on how to teach reading, and so I am suggesting here that this is something to worry about, particularly if the research that is submitted to you, or one must look at, comes from someone or disciples under such a given situation. I'd look particularly carefully at such research and worry, though I don't know the condition can be easily corrected.

Mann: There are also stronger expectancy effects working with handicapped than with normal ones, because there is so much hope riding on interventions—really—and I don't think it implies dishonesty, or intent to tamper, but it certainly affects the results you get.

French: Oh, yet...But, I don't think it is only in special education that you get that kind of a problem. I've seen it in a lot of places, and it is a particular strike against dissertation research. The guy who does the dissertation hypothesizes that something is going to happen and invariably it does. Too many dissertations have come out too favorably—then later replications just don't have them hold up. On this basis alone, I don't like to put a lot of faith in pooling data from a series of dissertations done in various settings. It may be a reasonable thing to try, but there are dangers.

Mann: Certainly the student generally wants to produce some positive results. Jay, how do you respond to this issue?

Millman: Well, you know, I think we are in a very sorry state of affairs if this is the case—that people, you know—people purposely do things in order to get results they want.

French: I don't know how conscious the fudging is, but have found that if the guy writing the dissertation is also the trainer in a program—the stimulator in a program—that he, in that instance, does a better job than when that program is not a part of his dissertation.

Millman: That doesn't bother me too much. What I thought you were saying is that there is fudging...

French: Well, I've read sometime ago that Greg Mudraker, in his autobiography, talks about how he changed from psychology to another field because...well, the graduate students were falsifying the data to please the professor.

Millman: Right. That's a mistake, and I don't think our reward system should bow to that. That is, I think journalists ought to print negative results as readily as positive, provided that there is a rationale for expecting the positive results. Some people claim that journals do not publish negative results—and that's a little bit tricky—and yet, when you don't see negative results in the journal, it may be that when you don't get positive results because the research wasn't very good to start with and your conceptualization wasn't good and the treatments weren't good and was it worth printing, it wasn't for that reason—not because the results came out negative. But, I think, if there is this fudging going on, that it is a sorry state of affairs and we must do all we can in our reward structure not to reinforce fudging. You'll have to limit dissertations—regardless of which way the data comes out, provided that the conceptualizations are adequate up to the time of the data—and let the chips fall where they may.

Mann: But, Jay, I would assure you—I wouldn't assure—I'm asking a question. Is the study, if done poorly and has negative results, are chances inclined less favorable for being published than if it is done poorly and provides positive results?

Millman: Well, I think that is too bad, but all things being equal, you know, they should be equally as acceptable as positive results. I'm not saying that is the case, but I would hope it would be.

Feldhusen: One way of introducing a slight corrective factor into all of this—not a total one—it would seem to me advantageous, in the special education domain, to find ways and journal editors I hope will help, to make data of studies easily available to those who would replicate or re-analyze the data. At the present time it is almost impossible to get such data. Cronbach and Snow seem to make quite a habit of doing this, too—trying to get data and re-analyzing, but always it is a terrible problem. You don't get the answers to the letters you write, and you wait, and wait, and wait. This would be one kind of an aid: to make the data available to others who would re-analyze it.

Millman: I could make an answer as to how to do that, of course, and that is to have it deposited in a national depository and what we did when I edited it was when an author sent me a manuscript which I felt, or a review committee feld, had too many tables—you know, it just wasn't worth all that journal space—we frequently had them make it available in documents elsewhere. However...the raw data?

Feldhusen: That is what I am suggesting...

Millman: Yeah...I was thinking it could be made available, and that might be one way. I guess I would agree with you...relying on an individual author to supply it is not very good. Two years later, that guy is no longer interested in the research or has destroyed his data. If data are important to make available to people, then we should not rely on the individual for data. It should be kept elsewhere!

Feldhusen: Well, one certainly concludes from the quite large number of re-analyses that Cronbach and Snow report doing in that monograph on ARI and where they claim that the data was incorrectly or inappropriately analyzed to begin with, that journal editors ought to be kind of worried about this, and that otherwise, people in general, ought to be on the spot in this way—to make their data handily, readily available to people who would like to re-analyze or re-examine. That data, of course, could still be wrong—could still be fudged up, but...

Mann: You would make a lot of uneasy people out there in the field call for data that might be re-analyzed.

French: It could all be microfiched, and in government projects at least, they could store it for you...put it in their X system somewhere, or whatever.

Mann: I had a recent incident, by the way. An author submitted an article. I wanted an additional fairly easy analysis from him, and I called him up and said, "Would you mind doing this?" And, he said, "Well, I threw the data out." It was fast. He didn't have the data available. So, we both laughed and I hung up and never heard about the article again. It does happen, you know. We authors pray you won't look too carefully, and ship it in and hope for publication, especially in a field where interesting studies are hard to come by and you are looking for something new, in terms of information. The novelty factor is critical, perhaps more in special education than in others, but maybe not...

Hopkins: I think it is a very good idea and I hope that eventually it becomes a condition for publication—to have the data submitted and available at some central source. I recently had occasion to go way back into research, in a given area, and almost all of it was useless because it was...I'm not sure whether it was correctly analyzed or, in those days they didn't do things the way we do, and so I came up empty-handed.

Mann: The raw data itself wasn't available, you are saying.

Hopkins: Yes. As John mentioned, it is very, very hard to get someone to let you have his data, even if he has it.

Mann: Is yours available, Rosa?

Hagin: I remember some document coding for the retaining of data for five years, and certainly then the next step would be this business of storing it. It would be really a very important thing.

Feldhusen: This has a tie-in with Jay's idea, too, and it suggests yet another way of this antidote for your small N problem in special education. If there is commonality among projects at all, and if data is deposited, we can also begin to look upon it as a potential source from which we could pool results of a number of studies of small N which, over the long haul, if people go at comparative problems, might provide us with a set of data that adds up to large N information that might be a partial antidote to the small N problem.

French: There might be the ethical problem of releasing critical data collected from a kid to somebody who wasn't involved in the original study.

Mann: Do you think this would be critical?

French: Well, I am sure there will be some critical groups, either right or left, who will feel that you shouldn't allow me to see the data you collected.

Hagin: If the subject were given you with confidentiality involved...

French: Yes, that is right. For example, if you talk about five kids of a certain kind at the State College of Pennsylvania, there may be certain characteristics that might be farily easily identified, even though...I think the names might have been removed.

Hagin: Well, this would move in a direction I have been looking forward to for a long time. Also, while we are at it, why not complain about the people who do research in vacuums, and many small investigators are guilty of this—they don't do their homework. I find that investigators are investigating problems that I thought were taken care of twenty years ago. There is, lo and behold, in this catalog of selected documents of the American Sociological Association, Volume 1, of the study of eyedness, handedness, and footedness in predicting success in college. Now this has been done, and done, and done, and summarized. And, lo and behold, it appears again in thirty cases, and cells of one and six. Now, if there were some kind of data bank and more data availalbe, it might not be necessary for this fellow, who is going to junior college, to worry about thirty cases—so many studies have been done.

Feldhusen: You might not only indict the man who did the research, but the editorial screening.

Mann: But, is this?...There is an awful naivety, I think, in most of us, when it comes to moving a little further in, of our own particular areas and I think you will find psychologists doing things that have been beaten to death in child neurology or whatever, for years—and in special education they discover that, lo and behold, it has been discarded in child development. Other eminent journals publish the stuff and it is very tired stuff, because it was discarded earlier and was not visible to the trained vision of the researcher and the field.

Proger: How about the research interpretation thing? Could we now consider it as a topic.

French: Well, one of my problems is determining to what is such a given sample generalizable to, when you do a study in a particular community and with these things...how do you know that community is like any other community?

Proger: I'm especially worried about how results of both basic and applied research will filter down to educational practice. We don't see them filtering down. I think we can problably agree on that.

French: I think it relates a lot to what we were saying about being able to pool and summarize data from various places, if their methods are reasonably similar. The, there is more opportunity to put them together and come to some conclusions—some generalizations. I think, particularly in the area of the gifted, the ability to tie one study to another is very, very difficult because the definitions of giftedness vary so tremendously and the instruments used to make selections vary so greatly, that it is very difficult to come up with conclusions about the field.

Mann: Well, this pooling concept, you know...can we return to it longer?. Is it really a legitimate goal, in terms of methodology? Can we indeed pool, over years and with different groups, these results from various studies?

Millman: I am not sure I have anything to say about that. I think people who try to do this are usually frustrated because the data that somebody else is studying just isn't exactly what I had in mind. But, I do think that we ought to perhaps make better use of, you know, information we gleaned from others. There is the Bayesian approach again which, in a sense, says look at all the previous stuff that you can get hold of on your particular topic and make a judgment about your expectancies and about how your studies are going to work out. Then, there are a formal set of procedures by which you can use those expectancies, in evaluating your present current study, and frankly, that is one aspect of the basic approach that I would take on making use of previous research—instead of pooling the data in some sort of systematic way. That becomes kind of frustrating, and the data just doesn't really lend itself to pooling. The replications, if there are indeed replications, aren't really identical enough to allow for that kind of pooling, and what I would best be left with is some general feel about the phenomenon.

Feldhusen: In spite of negative things that I said earlier about the indictments of some Gluck assessments, I think that it would be of considerable interest to me if I could get hold of, from a data bank, the results of all the different studies that had been done in the U.S. and even in other countries with the Gluck assessments data. Now delinquency criteria, to be sure, are slippery and difficult things, but nevertheless I have some faith that I could carry out analysis with such data that would offer me some advantages over the kind of sample that I worked with—my own sample. Certainly it would give depth and diversity to the child population—much greater than what I have in the particular sample I have. I have the sample from one county in the U.S. and a slightly unusual county in some ways. So, I think particularly and for prediction type studies, this happens to be the one that first comes to mind—that pooling of data across a variety of studies could possibly lead to quite interesting and useful analyses.

Millman: It certainly depends on the purpose of studies and the nature of the study. These are the kinds of prediction studies that are like those dealing with effectiveness of certain drugs, for example, where there are kinds of established indicators. These certainly lend themselves to pooling, whereas in other situations where you would make some sort of quantitative assessments or studies on the effectiveness of the open schools where data is vague, forget it.

Feldhusen: Well, you would inevitably be forced to make some judgments about the comparability of your predictors and your criterion data and would be operating at some hazard to be sure. I don't, for the moment, know of any empirical way I could arrive at comparisons among studies. Again, as we mentioned earlier, the more data or the more information each researcher supplies, when he deposited his data, as to what went on, what he did, what went right or wrong, how the treatments were administered, the more I would be in a position to draw out of the bank, to put data together for some type of analysis useful to me!

Hopkins: Well, I think it depends on the type of pooling we are talking about. I don't think there would be a mess if the investigator deliberately set out to kind of replicate someone else's study, for example, and so he had the same amount of time between pretests and post-tests, and all that. Pooling here would be dubious because of different type subjects, experimenters, locales and biases. It can be recommended in certain situations, however, One study that comes to mind was one going on at JFK, where they have a stimulation program for newborns, and it is going to be a controlled experiment situation but they can only treat a small number at any one time. While they are going to continue to let the study run, that is, babies will run through and be assessed at certain periods of time but they will hit different calendar times. Now, even though they can handle only a few at one time, over a course of a few months, they will be able to develop a very sizeable N. At the same time, they can test continuously to see in which direction they are going. So, I think, that kind of pooling is very promising...where your treatments are individually administered and hence, can begin the study at any time.

Mann: But not in terms of a group situation where it is so hard to duplicate the replicants.

Hopkins: That's right.

Proger: I don't see how, even if we could legitimately pool data with comparable studies from special education on EMR's, etc. with regard to certain goals in mind that each study was trying to accomplish one after another. I still don't see how in the world that is going to affect the actual functioning field. I don't see how data pools—they are nice and I like them, and I like Flanagan's ideas of the huge study in the sky, pooling everything—but I don't see how that is going to help the field. Can we consider the broader issue of how we translate research for the teacher?

French: The teacher really doesn't care about the research.

Feldhusen: Well, of course, that is documented now—that very few people read research articles. However, the model established by the R&D centers and the regional lab, seems to me to be an awfully good, viable and large-scale systematic way of translating research into practice. I am impressed that the R&D centers and the regional labs, right at this point in time, have caused large scale educational innovations statewide. For example, from the Wisconsin R&D center...what do they call their system?...IGE, or whatever it is, is being implemented in a wide scale in the state of Ohio, in a large number of schools. So that what began as quite basic research, has over a period of years, been finally brought to a point of widespread educational application. I guess the major message from it is that it won't happen just by chance and on a piecemeal basis, as a result of journal publications and individual teacher applications. It certainly appears that to make significant changes in educational practice in the schools, that a much larger engineering thing is necessary. That is a point, incidentally, that Richard Schutz argues also in a new paper from SWRL. He sees that as the only way that major changes will come about in educational practice.

Mann: Well, this might be a useful topic to conclude on. Is everyone in favor of the regional labs and R&D's approach? There is a danger of a vested eggs in one basket interest and promoting their products, and also, you are going to certainly have the investigator or the advocate effect creating results, and so on. And, is it really translation of research into practice?

Hopkins: Well, the lab approach, I think in certain instances, certainly has developed materials curricula that have been and are being widely disseminating and are having a direct impact on what is going on in the classroom, which I think addresses the point that Bart raised. However, I do not see that these materials have, in most cases, really been the result of basic research as I would think of it. I think they are curriculum development and I don't think there is that much data...is there, John?...in terms of how much better SWRL materials are than by not using them, in terms of that last step. I think they're criterion-referenced oriented but I think, in the final analysis, there has to be a comparative decision made in terms of what they are replacing. Is what they are replacing worse than what they are getting.

Proger: Yes, they are peddling their own brand of wares. They probably are very well thought out—perhaps have a hierarchy of this, that and the other thing in it—and that is good. But, I wonder...but still, I don't see that as research brought to bear helping the field.

Feldhusen: Again, I invite you to read the first publication in this series, which bears particularly on the SWRL development. Now that it is all done, or nearly done...now the group are trying to document the whole long course of how they finally arrived at this instructional program that can now be disseminated through many states, and of course, they are arguing strongly in trying to document, as I said this morning, their way back to their basic research that culminates, finally, in this thing that is being disseminated. Now, I do not know...this is certainly one thing...I don't know how we will ever provide empirical answers to your questions, but later argument is persuasive and the particularly interesting thing about this single document is that Dick Schutz is described as having been involved in the whole continuum from a long time ago. He is not a terribly old guy, but he claims to have been involved in basic research which all filters up, finally, to this eventual product.

Millman: One promise of the lab and R&D centers is that centralizing the large amount of these sources in one institution can permit the conduct of the programmatic work through a particular area. Whether or not this turns out to be what goes on—that is another matter. But, at least, that promise is there.

Bracht: I think that is the value. That there is a programmatic approach to doing research. We set a goal and we now have resources to conduct the studies necessary to achieve that goal. So, we identify this study as being important and then another, and another one...some going on simultaneously, which provide direction for future studies. That didn't happen before we had the regional labs and the R&D centers. Each person did his thing and sometimes several people would be interested in the same thing, but we did not have the intense programmatic effort to try to solve problems or achieve some goals. I think that is what we need, and have gotten, from the labs and R&D centers, to some degree.

French: So, I don't know a lot about it but there is one thing that I am concerned about. That is the profit motive in the developmental materials and competition of the R&D centers with the private sector, in the development of materials. For a long time now it has been Scott Forseman and Houghton-Mifflin and Addison-Wesley, and a bunch of book companies that have been developing these things and they have carried us a long way and then what happens when an R&D center plows millions of dollars into the development of something that is in a very small segment actually in the total educational program, then what happens to that? Is that in the private sector? Do the publishers bid for the opportunity to publish it or how is dissemination going to take place when we have that massive federal support and development? I know of federally funded projects whose authors are making mints out of selling the stuff, even though it is in the public domain...but I think that is something that the field has to look into.

## REFERENCES

Barber, T. X. Pitfalls in research: Nine investigator and experimenter effects. In R. M. W. Travers (Ed.) Second handbook of research on teaching. Chicago: Rand McNally, 1973. Pp. 382-404.

Begab, M. J. Some perspectives on research in mental retardation research centers. American Journal of Mental Deficiency, 1973, 77, 483-485.

Blackman, L. S. Research and the classroom: Mahomet and the mountain revisited. Exceptional Children, 1972, 39, 181-191.

Blatt, B., & Garfunkel, F. Teaching the mentally retarded. In R. M. W. Travers (Ed.) Second handbook of research on teaching. Chicago: Rand McNally, 1973, Pp. 632-656.

Brown, L. Instructional programs for trainable-level retarded students. In L. Mann & D. A. Sabatino (Eds.) The first review of special education. Philadelphia: JSE Press, 1973. Vol. 2, Pp. 103-136.

Chin, R., & Downey, L. Changing change: Innovating a discipline. In R. M. W. Travers (Ed.) Second handbook of research on teaching. Chicago: Rand McNally, 1973, Pp. 513-529.

Clifford, G. J. A history of the impact of research on teaching. In R. M. W. Travers (Ed.) Second handbook of research on teaching. Chicago: Rand McNally, 1973, Pp. 1-46.

Dershimer, R. A., & Iannaccone, L. Social and political influences on educational research. In R. M. W. Travers (Ed.) Second handbook of research on teaching. Chicago: Rand McNally, 1973. Pp. 113-121.

Drew, C. J., & Martinson, M. C. Educational methodology: An examination of approach. Exceptional Children, 1971, 38, 117-120.

Fink, A. H., & Glass, R. M. Contemporary issues in the education of the behaviorally disordered. In L. Mann & D. A. Sabatino (Eds.) The first review of special education. Philadelphia: JSE Press, 1973, Vol. 2, Pp. 137-160.

Frierson, E. C. The gifted. Review of Educational Research, 1969, 39, 25-37.

Games, P. A. Type IV errors revisited. Psychological Bulletin, 1973, 80, 304-307.

Getzels, J. W., & Dillon, J. T. The nature of giftedness and the education of the gifted. In R. M. W. Travers (Ed.) Second handbook *of research on teaching*. Chicago: Rand McNally, 1973, Pp. 689-731.

Glavin, J. P., & Quay, H. C. Behavior disorders. *Review of Educational Research*, 1969, 39, 83-102

Hewett, F. M., & Blake, P. R. Teaching the emotionally disturbed. In R. M. W. Travers (Ed.) *Second handbook of research on teaching*. Chicago: Rand McNally, 1973. Pp. 657-688.

Kass, C. E. Learning disabilities. *Review of Educational Research*, 1969, 39, 71-82.

Levin, J. R., & Marascuilo, L. A. Type IV errors and Games. *Psychological Bulletin*, 1973, 80, 308-309.

Lilly, M. S. The impact (or lack of it) of educational research on changes in educational practice. In L. Mann & D. A. Sabatino (Eds.) *The first review of special education*. Philadelphia: JSE Press, 1973, Vol. 2. Pp. 203-227.

McLaughlin, J. A., Hinojosa, V., & Trlica, J. Comprehension of statistical terms by special education students. *Exceptional Children*, 1973, 39, 408-412.

Miller, J. D. Request for data. *Journal of Speech and Hearing Research*, 1969, 12, 676.

Narin, F., & Garside, D. Journal relationships in special education. Exceptional Children, 1972, 38, 695-703.

Nolan, C. Y., & Ashcroft, S. C. The visually handicapped. *Review of Educational Research*, 1969, 39, 52-70.

Orlando, C. P. Review of the reading research in special education. In L. Mann & D. A. Sabatino (Eds.) *The first review of special education*. Philadelphia: JSE Press, 1973, Vol. 1, Pp. 261-283.

Prehm, H. J., & Crosson, J. E. The mentally retarded. *Review of Educational Research*, 1969, 39, 5-24.

Proger, B. B., Haughey, C. R., Spaans, D. N., & Proger, H. M. Large-scale, personalized information retrieval of psychological and educational research findings for school district decision making. In A. D. Berton (Ed.) *The social impact of information retrieval: The information bazaar; Seventh annual national colloquium on information retrieval*. Philadelphia: Medical Documentation Service, The College of Physicians of Philadelphia, 1970. Pp. 272-296.

Proger, B. B. The fine art of educational research muckraking. *American Educational Research Journal*, 1971, 8, 163-166.

Quigley, S. P. The deaf and the hard of hearing. *Review of Educational Research*, 1969, 39, 103-123.

Rutter, M., & Yule, W. Specific reading retardation. In L. Mann & D. A. Sabatino (Eds.) The first review of special education. Philadelphia: JSE Press, 1973. Vol. 2, Pp. 1-50.

Short, E. C. Knowledge production and utilization in curriculum: A special case of the general phenomenon. Review of Educational Research, 1973, 43, 237-301.

Shulman, L. S. Reconstruction of educational research. Review of Educational Research, 1970, 40, 371-396.

Sitko, M. C., & Semmel, M. I. Language and language behavior of the mentally retarded. In L. Mann & D. A. Sabatino (Eds.) The first review of special education. Philadelphia: JSE Press, 1973, Vol 1, Pp. 203-259.

Snow, R. E. Theory construction for research on teaching. In R. M. W. Travers (Ed.) Second handbook of research on teaching. Chicago: Rand McNally, 1973. Pp. 77-112.

Starr, C. The speech handicapped. Review of Educational Research, 1969, 39, 38-51.

Travers, R. M. W. Educational technology and related research viewed as a political force. In R. M. W. Travers (Ed.) Second handbook of research on teaching. Chicago: Rand McNally, 1973. Pp. 979-996.

Weiss, D. J. Test reviewing: Problems and prospects. Minneapolis, Minn.: Inter-Association Council on Test Reviewing (University of Minnesota), 1970. Available as ERIC document ED-044-439.

Worthen, B. R., & Roaden, A. L. An analysis of the research involvement and productivity of AERA members. Journal of Educational Research, 1972, 66, 39-45.

## CHAPTER 14

# THE COMPLEMENTARY ROLES OF DESCRIPTIVE AND INFERENTIAL STATISTICS

Overview: In formal research endeavors of education, descriptive statistics have been relegated to a minor, and in some cases nonexistent, role. Most researchers have tended toward the seemingly more sophisticated inferential statistics model. Chapter 14 puts the two types of statistical procedures in proper focus and lends many novel and useful applications of each area. In formal research or in applied program evaluation, observational techniques can be the primary technique of generating conclusions in a descriptive, noninferential manner. However, observation can also be used to complete the inferential statistics of a research study in which control problems have arisen so that qualifications on interpretation are necessary. Apart from research usage, however, observational techniques can be of immense value in practical decision making. First, one can view observation in a group-descriptive sense as a type of formative feedback to an administrator to allow him to change a foundering program. Second, one can use observation as a diagnostic appraisal method to study intensively the problems of an individual child and then use such feedback for diagnostic-prescriptive teaching. Thus, this chapter breathes new life and vitality back into what has usually been a topic treated with condescension.

Proger: Leave the ATI business alone for the time being. What role do you see case studies and/or observational work playing in identifying some of the kids' problems? How does informal observation play a role or should it play a role in conjunction with formal diagnosis via testing? What is the role of case studies and observational work in special education, or in education in general—in educational research in general, if you wish? What should be the role?

Mann: I wonder if I could briefly raise the issue which I saw in the AERA newsletter about descriptive statistics being perhaps the instrument of choice, as opposed to inferential statistics, for a lot of studies. It may have some implications for special education. It referred to the national assessment work as being a far more valid path of measurement than constantly seeking out inferential, statistically significant differences.

Bracht: I think there are situations in which descriptive statistics are appropriate, and other situations where inferential statistics must be used. I think it depends upon the purpose of the study and what the researcher intends to do with the results obtained. If we are carrying on an evaluation of an educational program, and we want to conduct primarily a formative evaluation—we want to see how a program is working in that situation and find ways of improving it—inferential statistics, I don't think, have any value. There, it is solely a use of descriptive statistics. If, on the other hand, we are comparing alternative treatments, and intend to make some decisions about the usefulness or the comparative value of those two treatments within some population, inferential statistics are necessary. This might generate some disagreement, but in general, I think it is difficult to ask should we use descriptive or inferential statistics.

Mann: I think some concern was expressed, in the article I read, about abuses of the inferences generated.

Stanley: Les, there is a special problem here that keeps coming up. I used to hear it years ago at Wisconsin, when one old-timer taught statistics. It goes something like this. Suppose that I grabbed the first fifty people that walked by that door. Suppose I do an experiment with them—a controlled, randomized experiment. Can I make inferences? Well...yes and no. Obviously, I can't treat them as a random sample of any defined operational sample population, but I can use an internal randomization model and get a randomization test due to the hypothesis for that group. One method works better than another method. Then, if I can argue logically that there probably aren't characteristics of that group that interact with the treatment variable to give a peculiar result compared with some other group I might be interested in, I can probably extend my inference to other groups, not probablistically, but sort of logically empirically. Now, a lot of people don't seem to realize that it is perfectly feasible to take those fifty people, assign twenty-five of them randomly to method A, and twenty-five of them to method B, do an experiment—use a randomization-permutation type test—and get a significant statement of a P value...Say some P value characterizes that particular experiment. Well, that is kind of a bridge between the purely descriptive and the real linear model of analysis of variance which requires random sampling of experimental units from an inferent population.

Hopkins: Just another point there. Aren't the results, Julian, when you do use the randomization procedure, essentially those you would have gotten from a classical inferential?

Stanley: From a T-test, yes. The randomization procedure which Fisher pointed out years ago is, for that situation, the definitive procedure. It is very well approximated by an F or a T test.

Hopkins: I don't think I would go quite as far as Glenn suggested. If you have interest in knowing whether the boys in the study have better writing ability than the girls, and if there are twenty-five of each, then it is of some value to know whether their differences are something that could have easily occurred by chance or not...even though there is no inferential issue, in the usual sense.

Bracht: I guess, thinking about the usual situations in which I think of doing formative evaluation, I think of other problems of interpreting. For example, the females, maybe in that situation, are not representative of males and females in the population. I guess I think there may be cases where truly inferential statistics might be appropriate, but I think, in most situations, we are just working within some context and we are really not investigating how males and females differ on some outcome or some trait. We are just interested in how they compare within that study itself, in a descriptive sense.

Stanley: You could still use the randomization model for this, too. It would work the same way, generally.

Feldhusen: I don't see a particular need to limit Glenn's comments to a formative evaluation. I see no reason why descriptive statistics are not entirely appropriate for a summative evaluation of the entire population in a school system who will receive a treatment of some kind or some new instruction, or what have you. If all the children have got it, as it were, why not? That certainly seems a classic case for descriptive results, doesn't it?

Hopkins: I see a problem in that. I think even there, you are interested in whether the results would replicate over the next year's students in the same school. If not, the real question is whether it should have been evaluated anyway, because it is reversible.

Feldhusen: It seems to me that some problems will arise. The emphasis in current evaluation is on the supplying of information to a decision maker, which clearly suggests what you are implying here, whereas really the old condition that Stufflebeam decries is the evaluation that is done, delivered—sort of end of the story—a description of what happened. That clearly is the case for descriptive analysis, but indeed, since this is implied in the CIPP model that we are supplying information to the decision maker, then your point is well taken...Then an inference is being made.

Mann: Jay?

Millman: That article that you read, Les, is one of the few articles that so incensed me that I was compelled to write a letter to the editor, which was subsequently published in the next issue. My response to that article, very briefly, went something like this. Inferential statistics are almost always of interest, and should be looked at more frequently. True, there have been many abuses to inferential tests, including significance tests, but there is a very important role that significance tests play and their abuses should not force us into the position of forgetting this particular role. Rarely are we interested in Mary Smith or Tom Jones, who may have been the subject of our particular investigation, as Ken quite frankly said is often done. What we really, either implicitly or explicitly, are saying is that we are really interested in kids like this. This is where inferential statistics plays a part. It helps us decide how much wiggle there may be in the descriptive data that we see on Mary Smith and Tom Jones. That is, it helps us to evaluate whether the observed findings, the observed differences and so on, may have been caused by chance alone. It helps us to assess the contribution that sampling error may have made to the findings, and in that sense, serves as the way of evaluating sampling error as a source of internal validity. You can use that kind of jargon—you probably shouldn't. It helps us to evaluate sampling error as a bogus reason for some findings that otherwise might be taken as real differences.

Bracht: One problem is that some researchers have tended to put a higher value on inferential statistics than descriptive statistics. I had the impression that it is somehow better to use inferential statistics...not in specific situations, but somehow—generally. This is what has led to some of the abuses Jay refers to. I am speculating that such a general impression does exist. If we can remove that feeling, then I think we might make some gains and make more appropriate use of different types of statistics. Another related point is that I think we have seen some suggestions recently that there doesn't always have to be an empirical comparison made in evaluation studies. John mentioned the first day, I think, we can use absolute standards of some sort. We expect performers to be at this level, and we can compare the performance of this group then with that standard, and there we do not have the tools for inferential statistics to help us. Criterion-referenced measurement is a case in point. I think that sort of thing is good. I don't want that to replace use of a control group in all situations, but I think it is another thing to include in our repertoire, to help us make decisions.

Feldhusen: Generally, one can't put much confidence in the results of a study in which differential, contaminating biases were operating that were uncontrolled. I would agree with Ken if the report of the study tells us only that there was a teacher who administered certain generally described treatments. But I do see one ray of hope that brings the study that you are otherwise ready to consign to the zero heap... I see one way of bringing it back to usefulness and that, again, is back to this route that Terry Denny has suggested and that is if you can possibly get somebody to get as good a description as possible of how the teacher behaved or operated—even if it is just one teacher. If I can have specifications somewhere or other of what she did, then I can at least begin to make some inferences about how other teachers operating, hopefully, in exactly the same way—now, again, this is N of one which is hazardous, but it is better than knowing nothing, and I would still hang on to the study if I had some information about the teacher in the class.

Hopkins: I think, in that situation, I think you've convinced me if you view it as kind of a case study strategy and present it as that, rather than Miss Smith doing this and Miss Jones doing that and they have a significant difference. So, a lot of people reading it say, oh, it is the materials, rather than Miss Smith. They don't realize that—or in your approach which is more of a case study, I think—there is very little danger of misinterpretation, or at least it is much less.

Feldhusen: And incidentally, that...well, you know, that is a fairly recent...that is, that issue. I am just reading it from that recent review on education evaluation, and the review of that research, and it does so happen that, according to their review, that there are an awful lot of skills and systems now for observing, or for getting information, about what is going on in the classroom. Just as an aside, we ought not to be totally despondent about getting a representation of what went on in the classroom during the research.

Hopkins: Isn't it also true that a lot of your treatments are more individually dispensed, like your perceptual training treatments that, probably, I don't know—do you work out a regimen that the kid goes through at home and it is less teacher-dependent, or would that be a poor example?

Hagin: Materials are written out and teachers are supervised as they do this. They keep a diary. There are several kinds of checks, but they do want their instructors.

Hopkins: They work with a whole group.

Hagin: No, one child at a time.

Hopkins: So to the extent that this is true, things are probably better than they are with normals.

Proger: You are sort of getting back in what we started at the very beginning this morning. The idea of prescriptive teaching where, you know, you have a separate treatment in effect for each kid although you could classify them under some broad umbrella and say prescriptive limited such and such techniques in general. You still find that every kid is a different treatment—in effect.

Mann: You can certainly handle the delivery of services far more effectively as a variable if you keep their modifications of material to a minimum.

Hopkins: I worked with a group of occupational therapists doing research earlier in the year, and almost all of their treatments are, you know, individual brushing of kids and things like that, which is very nice, in that we do not have to worry about this awful thing of the experimental unit as much as we do in a lot of other instructional research.

Bracht: If we use observational techniques to report what occurs during the treatment, do we use that in any systematic way in the analysis? Or, is that a kind of analysis where, prior to our analysis of the dependent variable data, we first look at the results of the observation data from treatment A and treatment B and see if they are similar. If they are similar, then go ahead and analyze the dependent variable data, or is there some way we can systematically integrate these into one analysis?

Mann: Deep silence over here...Ken?

Hopkins: Well, I suppose if you use the case-study approach, you've got all kinds of elbowroom because, hopefully, you just described what you did. And I think you are talking about a more credible establishment of evidence that you would get from a case-study approach.

Bracht: Yes, I guess so. Using some systematic observation techniques for a week and whether it is a system like Flanders or some other system where we record what is going on during the treatment.

Mann: Are you saying, wait until the study is over and then see if this is important, or just go into the design.

Bracht: No. I am asking how do we use that data? If we have done this, how do we use that data in the analysis?...A kind of separate analysis, or is it?...

Millman: I suppose a Bayesian would say that you, you know, that you would use that to adjust your prior probabilities that would then affect how the hard data are looked at. Traditionalists would probably say, do your straightforward, statistical analysis with the data you got, and then put your interpretation and judgment on at that point; this other information comes in handy for such qualifications. I guess you just make your judgment to be either early or late, as to when the observational data enters the analytical picture.

Feldhusen: I am not sure I understood you. Suppose I have an assessment say, of the mode of operation of the teacher. Let's pick up Flanders and find that a lot of my teachers are highly direct and others are indirect in their teaching behavior. Are you saying that if my sampling unit is the classroom and/or the teacher, then I might use indirect vs. direct as a classification variable for teacher, along with some other treatments that she may be trying to administer?

Millman: I wasn't thinking along those lines. Perhaps I misunderstood Glenn's question. I wasn't thinking of using the observation data as a classification. I was just saying, you would run an experiment in the usual sense and also have some other information you now know about the experiment. Let's say, the quality of the teaching or some very big case study, where you have nonquantifiable kinds of information—how do you make use of that descriptive data together with the hard, statistical stuff? I'm saying that you have a choice. You can do the hard statistical stuff and then you go to interpret it, you keep in the back of your mind all this anecdotal stuff that came along, which is the traditional way.

Feldhusen: Are you rejecting my notion, then, or would you accept that, too? I just want to be sure.

Millman: No.

Mann: Well, what is the other way? Adjust your probabilities, you are saying?

Millman: Yes, and then do your statistical analysis.

**Mann:** On what basis would you adjust your probabilities?

**Millman:** Well, I am not a good spokesman for that point of view, but I guess you would have to say, knowing now this anecdotal information—probabilities getting certain kinds of results—I am not sure. You know, I am not that familiar with it to work it through.

**Mann:** How does it strike you?

**Hopkins:** Well, like Jay, I'm also not a spokesman for the Bayesian approach because I think...I can't feel comfortable with the basis that we have to set those prior probabilities, so I wouldn't try to defend it at all.

**Feldhusen:** Well, use my solution then, I guess, or else just cover the data.

**Hopkins:** Well, I think what you are suggesting, John, makes good sense to me.

**Millman:** Right...I, too. But, I was taking the situation in hand—that kind of information where you just get the once-over observational data.

**Mann:** Suppose teacher A—as the stormy—teacher C is the better teacher, etc. How do you manage this?

**Feldhusen:** You bring up a different, cruder level of observation. Most clearly, the structure I opt for is some kind of systematic way of quantifying and subgrouping teachers.

**Hagin:** Should that not be built into the design initially—if possible?

**Feldhusen:** Well, yes. I, of course, would like to see that thought about, from the beginning, and not something that turns up worrisome later. It would be hard—yet the best thing one could tell a designer at the present time is to conduct a needs assignment of what might go wrong in the study. In terms of descriptive research methods, it is a monumental duty—putting together all the information about the vast numbers of systems for observing in the classroom, to get a fix on what is going on there. One such effort is the Mirrors on Behavior, which I am sure many of you know. There are things to turn to. I am not able to say how good it is, but certainly the Flanders system, for example, has been researched a vast amount. It is also true, in relation to your particular situation of special education diagnostic-prescriptive teaching, especially designed for observing one-to-one—tutorial-type interaction—are available. There are some people that have considerable faith that there can be great variation in what goes on in such situations and thus, there may be a need to make an assessment. I have just worked on that problem myself with tutors in a study, and have developed a crude observation system. I am convinced that, indeed, there are horrendous differences between tutors who seemingly had very precisely specified duties, precisely specified material, and who yet seem to depart markedly from time to time from the specifications.

French: Well, when the study is done, some might be in one cell and some might be in another, but you would not have predicted that at the time the study began.

Feldhusen: I must add that, for the present, I am operating merely with a sort of case-study observation approach. I have no fix within the study. My *N* is so small that I am simply finding it out and it is going to qualify the kind of conclusions or generalizations I draw. I am not able to use the information I've got systematically within the design analysis. I am thinking of the awful plight of observational study of the type you are talking about where one might end up as left with 400 variables and what to do then. This...of course there are obviously answers as to what to do then, but...

Mann: I have recommended several. I can't use it.

Feldhusen: I don't have an answer at all. I am simply dismayed, at the moment, at how to draw together from an observational kind of study—the type you suggest—something that would still let me hang on to a characterization of the individual. I am quite ready, again, to apply techniques that will provide me with a way of reducing my idea so that I will end up with subgroups, clusters or descriptions of factors, or what have you, but I don't know what to make of an observational study if I am suddenly innundated with a mass of data to which I am to react clinically, individually. And yet, I guess, as Ken suggested earlier, out of such observation certainly brilliant insights occurred in the work of Piaget and others, but that, I see, as a very intuitive, clinical kind of thing and maybe that is the only answer for the moment. It does seem a problem...to me.

French: Is the value to be derived from intensive observational study worth the price of collecting a large enough sample of people with a particular characteristic in which one can study them as a group and the entire environmental interaction? That is, in many communities there are ten or twelve kids with a particular characteristic, but that isn't enough. Should we gather all the kids from many communities whose parents will permit placing them in an institution where we can at least study their life-style?

Feldhusen: We have a rule of thumb, in a way, that for any kind of multivariate analysis: the number of subjects has to exceed the number of characteristics we are going to deal with, and so we are limited. Take exception with that rule of thumb, if you like, but I think I understand what I am saying.

French: So, one answer might be to gather larger and larger groups of these kids together and place them in an institution where we can monitor their behavior on the variables we choose to study, for a particular period of time. Is that approach socially acceptable?

Mann: It has been done in New York State. Clausen Institute up there-they group them in a large institutional population. They just fly in certain kinds of youngsters for physiological study.

French: I think various places have tried that for years and years. A number of private institutions around here have made attempts in that direction, but there is always the social issue of whether or not you can...it ought to be done...take a child from his home when the parents don't-whenever the parents can't handle him.

Feldhusen: The city of Indianapolis has just established a day-type diagnostic center for children with learning problems in the school. So, a child might be going anywhere from the city, to come for a period of anywhere from one week to three or four months, for intensive observation. This is not--they have tried to create a continuing school-type of atmosphere there--to be thrown into the new group of children. Essentially, then, they try to turn on massive observational forces; first, to detect his problems, and secondly, trying to capitalize upon those observations to develop a regimen which they first want to try out right there. Then, hopefully, send him back to this school with that prescription, hoping that his teacher back there will be able to continue that regimen.

Bracht: There are limitations in resources available in what we can do and I think that we might start making school observations intensely-those who are not performing successfully with the program we presently have. This is taking the approach that we... there are certain outcomes that we want to achieve. Why should we spend our limited resources in studying those who are achieving those outcomes? Let's use our resources to study those who are not achieving those outcomes.

Mann: Wouldn't you say, for example, if you have certain retarded youngsters who are achieving well for their mental level, let's say reading, and some of them do, by the way; there is no doubt about it. Wouldn't it be worthwhile studying this kind of youngster, who is achieving, as compared to the MR the same ability level and associate mental structure, who isn't achieving?

**Bracht:** I think so, although I guess I'm putting a priority in that if we do not have enough resources to study both, that I would put a priority on studying those who are not presently achieving.

**French:** I think that is really what special education is—that is an over-simplification—but, those mentally retarded kids who aren't too deviant are still in the public school and we seldom get them involved in studies, even as controls—in one way or another. The problem is that there just aren't people who know how to observe, and there aren't observational scales that they could apply if they could be taught. It seems like this is a tremendous lag, but for years now, they have taken the kids who don't learn and put them someplace else and then what happens to them?

**Mann:** I think the issue of the number of variables needed to describe handicapped children in research settings is similar to the number of descriptors needed in information gathering, and when you want to go to the computer for literature searches with descriptors which are available, of course, you come up with a rather poor yield at the present time from the ERIC system, or anywhere else, or from manually conducted literature searches. The hand searching that we do is still more effective because we are bringing more descriptors or variables with us to bear upon the search. The same with special education or general ed; we have very few description variables to identify our population. Maybe the critical observation you are talking about would help here. We could pick them up intuitively or whatever, as Piaget did, indeed. Rosa?

**Hagin:** Another source of information we have found, fortunately interesting, is not isolating the children, or not institutionalizing them, because it seems if we believe Spitz and others this creates another variable to deal with. We have had the opportunity to see an intact group in a medium size school two blocks from the medical center, so it is really part of our community. We have seen every first grader for two years and we've continued to follow them so that the first group we saw is now in third grade. We have been able to use the clinical personnel to do the initial observations. In return for this, we have had to do some intervention, and in some ways we have spoiled the natural situation. We do have retrospective data on the children who are now fourth graders before intervention began. This has been a very useful teaching situation for our unit.

French: I guess, on the institutional business, I was concerned about the full day, and not the kind of observation that took place in the...in many institutions today. But, what happens when a kid wakes up at night and needs to go to the bathroom? You know, what kind of emotional support does he want, and does he get? Does this have...is this an influencing factor on how the kids learn in school? Rick Heeber's recent study about providing educators for infants whose mothers are below 70 IQ, or below 80 IQ, as a step in providing some educational intervention...but, it is still—and, I guess, it is some social intervention, as well—the type of people that are put in. But, my suggestions for an institution would be for that complete 24-hour day study, and not just education, but whatever is necessary in a social and emotional development of an individual, as well.

## REFERENCES

Birnbaum, M. H. The devil rides again: Correlation as an index of fit. *Psychological Bulletin*, 1973, 79, 239-242.

Darlington, R. B. Comparing two groups by simple graphs. *Psychological Bulletin*, 1973, 79, 110-116.

Ekehammar, B., & Magmesson, D. Intertrait correlations as a function of amount of judgmental information about the ratee. *Perceptual and Motor Skills*, 1973, 36, 373-374.

Gocka, E. F. Population estimates from nonvariant samples, with an application taken from the drug abuse area. *Perceptual and Motor Skills*, 1973, 36, 1207-1210.

Gordon, I. J., & Jester, R. E. Techniques of observing teaching in early childhood and outcomes of particular procedures. In R. M. W. Travers (Ed.) *Second handbook of research on teaching*. Chicago: Rand McNally, 1973, pp. 184-217.

Gorman, B. S., & Fleishman, A. I. Semitabular methods for calculating Spearman's rho. *Perceptual and Motor Skills*, 1973, 37, 30.

Harmatz, M. G. Observational study of ward staff behavior. *Exceptional Children*, 1973, 39, 554-558.

Hintzman, D. L. On testing the independence of associations. *Psychological Review*, 197e, 79, 261-264.

Knafle, J. D. The relationship of behavior ratings to grades earned by female high school students. *Journal of Educational Research*, 1972, 66, 106-110.

Koran, M. L., Snow, R. E., & McDonald, F. J. Teacher aptitude and observational learning of a teaching skill. *Journal of Educational* Psychology, 1971, 62, 219-228.

Medley, D. M., & Mitzel, H. E. Measuring classroom behavior by systematic observation. In N. L. Gage (Ed.) *Handbook of research on teaching*. Chicago: Rand McNally, 1963, pp. 247-328.

Quirk, T. J., Steen, M. T., & Lipe, D. Development of the program for learning in accordance with needs teacher observation scale: A teacher observation scale for individualized instruction. *Journal of Educational Psychology*, 1971, 62, 188-200.

Renzulli, J. S., & Hartman, R. K. Scale for rating behavioral characteristics of superior students. Exceptional Children, 1971, 38, 243-248.

Richter, J., & Scandrette, O. Relationship between teacher personality ratings and student academic success. Journal of Educational Research, 1971, 65, 127-131.

Rosenshine, B., & Furst, N. The use of direct observation to study teaching. In R. M. W. Travers (Ed.) Second handbook of research on teaching. Chicago: Rand McNally, 1971, pp. 122-183.

Shaffer, J. P. Defining and testing hypotheses in multi-dimensional contingency tables. Psychological Bulletin, 1972, 79, 127-141.

Smith, J. E. K. On tests of quasi-independence in psychological research. Psychological Bulletin, 1973, 80, 329-333.

Vergason, G. A. Accountability in special education. Exceptional Children, 1973, 39, 367-373.

Yairi, E. Item analysis for parental behaviors rated by stutterers and nonstutterers. Perceptual and Motor Skills, 1973, 36, 451-452.

## CHAPTER 15

# IMPLEMENTING A RESEARCH DESIGN WITH INTERNAL CONTROLS SUCH AS ADEQUATE SAMPLING, RANDOMIZATION, AND BLOCKING

Overview: The issue is raised as to whether or not the extreme variability typically found in any sample of exceptional children can be controlled adequately by the classical techniques of random assignment of treatments and blocking. A complicating factor is the frequent dearth of children available for participation in a study. The participants suggest several ways to deal with these problems. First, they question whether small Ns are really a necessary condition if adequate pre-experimental planning is employed to allow for widespread sampling to build up greater Ns. Further, small Ns may be an artifiact of the overly precise or narrow-minded thinking when defining their sampling parameters; by loosening the stringency of IQ band widths, for example, adequate Ns can easily be generated where otherwise one would be in trouble. Also replication of procedures with different teachers and in different settings, when built in as a control factor, will build up the N of a study. Second, one could collect various types of pre-data to allow later adjustment of scores if he suspects something went awry either during sampling or later during the course of the study. Third, if one does suspect that in spite of best experimental procedures, lack of control still resulted in certain sections of the study, one could have extra Ss in reserve to re-do, as it were, certain parts of the study. Finally, in preference to sometimes inappropriate data adjustment when one suspects that something went wrong in the control of the study, it is sometimes more desirable to block or stratify more precisely than one might do ordinarily even if cell sizes become quite small. In essence, this chapter seems to suggest that while there may be peculiar features of special education research (e.g., greater variability within subpopulations, small Ns), there is nothing that just a greater dose of classical design cannot cure.

Proger: One issue that particularly concerns special education researchers' experimental design is the control of extraneous factors. Let's take the extreme variability of the EMRs. How can one build into his design appropriate controls for contaminating influences? What are the roles of such techniques as randomization and blocking? Are they different from regular education design theory? Could we talk about the variability of retardates as in the city of Philadelphia? Suppose we are going to pull out a sample of those EMRs. We can define it by a certain IQ range, moderately educable retarded or mildly so, etc. However, within a certain range you still have a tremendous amount of fluctuation and there is marked inprecision in that labeling. Just when you think you have gotten a decent handle on the sample, you are fouled up by contamination within each stratum, such as emotional disturbance, severe cultural deprivation, etc., do you then block findings? Just when there is still so much variability there in spite of stratification. Now, do we use simple randomization, random assignment, or do we go into matching, as a means of accounting for some of that severe fluctuation that is left in the sample, even after precise stratification of the original population. I fear that, in contrast to research with normal children, research on even a very precisely defined sample of children may have so much "NOISE" in it that treatments "SIGNALS" are obliterated.

Cross: Well, I don't see that. I don't see that as an awful lot different than working with any other population. Anywhere there are selected groups you have variation. I don't see that it is a special issue.

Proger: But, I think I am implicitly assuming these kids are even multiply handicapped. You know when you say you have an EMR that you don't have a "pure" EMR. The fact that the kid is handicapped means he probably has many, many problems.

Cross: I think it still is a simple matter of hoping that by randomization you will take out the extraneous variables prior to treatment.

Feldhusen: Except that you may not be satisfied to simply, as it were, sweep them under the carpet. By randomization you may indeed be saying to us, "We want to know more about how this wide range of emotional stability or IQ or whatever might be related to the outcome variable." We sort of made the full circle back to ATI now. I'm beginning to hear it. And so, again, the randomization now simply creates the situation in which you hope that that particular thing, emotional stability, or IQ, or anxiety, or whatever it is, will not bias up comparisons among treatments. However, if you would prefer to get greater precision in your understanding of what happens with such variability, then you might choose some alternative route of trying to, perhaps, bring each, or some, of those things into your experiment as in what is called the classification variable.

**Millman:** you are now talking about blocking.

**Feldhusen:** Right.

**Proger:** I'm thinking of a small number of subjects, too. I truly doubt the effectiveness of randomization in controlling extraneous variance when you have very small samples.

**Mann:** Could we start this as a formal topic, perhaps right now, and discuss the whole issue of randomization, and go through specific steps or problems that face special education research, though they certainly remain in all of education. Now, randomization assumes, by and large, that we can handle the extraneous variabilities of our subject and with regular children—normal children—we can do it fairly safely. While a handicapped group can be identified categorically as being such and such, we can never really be sure that we effectively randomized this kind of a population into groups for experimentation and research. You can do it on an arbitrary criterion, but can you do it legitimately, in terms of the other pre-established problems that are there? I doubt it.

**French:** It seems like theoretically you can, if you have a large enough N, but we never do. What we were saying a little bit earlier was that we need to have a lot more descriptive characteristics of individuals with particular relevance to the tasks to be taught and that we were concerned with, about how big a sample one would need to define the miniscule parts of the learning task.

**Mann:** A critical issue—that most research with the handicapped is done with very small samples, and yet the variations are enormous. Now, is this an unsolvable dilemma?

**French:** Well, it seems like that indicates a priority for research in the area that is for the development of tools to assist with task analysis, so we can get in and explore the smaller components of each of the tasks that are in the big picture.

**Millman:** I think there is a definite answer to this kind of question that any of us can provide—Ken, or Glenn, or me.

**Hopkins:** Well, there are two related concerns, I hear. One is strictly a matter of efficiency, or power; that is, purely and simply, the variability within the group gives you a lot of knowledge of the system and not very much statistical power. To solve that problem isn't difficult as you stratify the group, and you can take a lot of that out. However, the tougher part is whether or not there is any action. That is differential treatment efficacy, depending on subpopulations that are only identified within that group, and that is a much tougher question, but the answer is simple; the solution may not be, in other words.

French: But to covary, you've got to have something quantifiable to use and I think the instrumentation in special education is so poor that, at this stage of development, that solution really doesn't work, except in theory.

Proger: ...with the assumptions behind it.

Hopkins: Well, it depends on variability of environments. If it is a learning treatment, in other words, it is simple enough to get a lot of information on variables.

French: ...by constructing new instruments for the studies.

Hopkins: By individual intelligence tests, or whatever.

French: No. The individual intelligence test just won't work. They are not precise enough to utilize in that kind of a situation.

Hopkins: Even if they are able to correlate?

French: Well, they can, but it is so gross that it isn't going to get at the specific task that the kids are going to be learning.

Hopkins: Well, I think that if they are just going to talk about this problem in the abstract, it depends upon the particular dependent variable in question—what the treatment is—whether or not you can get useful data as pre-treatment indicators. Some are tougher than others.

French: Intelligence is an important variable, but when you use it on a restricted range for a group of educables—three or four years age span—I think, practically, it is useless.

Hopkins: Then, the question is, "What is this variable for which there is so much variation among the group?"

French: The previous learning they bring to the task would be what I would conclude.

Hopkins: Well certainly, that should be capable of being measured, and then you could use that.

French: But not with existing instruments that are for sale.

Hopkins: That's a whole different issue.

Proger: I see this as sort of leading us into the fact of small Ns, too. John, in the study you submitted recently to the

Journal of Special Education on aggressive children, I think that had quite a few bodies in it, did it not? So, we were not too hung up by this, here. But, I am thinking of the usual type of special education experiment. Some researcher comes in and runs off 30 bodies. He uses what was available in the local institution of such and such a range of EMRs, and forms 3 different treatment groups. He uses randomization and theoretically, according to all the authorities, we have comparable groups. I wonder, really, how well randomization works with small Ns. My own feeling is that it falls down miserably. If you believe in the universal applicability of randomization, you probably don't see my skepticism.

Hopkins: Surely, your argument isn't with random assignment. The question is that there is going to be a lot of likelihood for type 2 error creeping into the situation. When you randomly assign subjects to treatment, you know any difference is random difference. You chance to see whether differences you observe are within the realm of chance differences or not. So, I don't think your concern is whether the statistical theory is really appropriate or not. It is whether you have enough efficiency in your research strategy to identify real differences. If there isn't, you have tremendous problems. That means you've got to have very potent treatments, using every conceivable apparatus at your disposal, to maximize power. Well, there are 4 or 5 paths of alternatives to increase power, but even if you are really limited and you rely on the failure to reject an old hypothesis as confirmation of your hypothesis, as most people do, you are in trouble.

Bracht: I'd like to bring up an issue that relates to this, and that is, do we, perhaps sometimes, put too much emphasis on statistical significance? Are we getting back to our point of this morning, and that is, quite often the findings that we obtain are context-bound. We can't generalize, anyway, to some large population because the findings obtained in one study relates very strongly to the context and we start study with them—the teachers involved—or advocates of the program in that study, or other factors. And so, we suddenly just lose faith in being able to generalize to some large population. If that is the case, how important is it that we obtain statistical significance? I just want to throw it out. It seems...

Millman: I think it is a very important one. On the scale from 1 to 10, I'd say its importance is 8.7. The reason I can say that is because if you cannot rule out the chance that is accounting for the differences, at least in that situation, I think you ought to know it; I think it is important. It is important to know whether in that context, and what was done—whether or not the treatments, or whatever—made a difference in that context, and I think that is important to know. On the other hand, I could not agree more—on the scale of 1 to 5 if agreement is 5—with your caution that you are suggesting that we have. That just because you can rule it out within other contexts, nor—which you did not say but I'm sure I would say—nor that that particular difference in magnitude was impractical. So, it is not sufficient to get significance to prove that it is going to be useful in other concepts.

Bracht: Necessary, but not sufficient.

Millman: Well, I'm not even sure it is necessary. I mean, if you have such small Ns and you don't reach significance, still you might reach practical significance if not statistical. And so, if you say that some treatment is not of any use, that gave practical significance, rather than statistical, because you didn't have enough power, that is, to blame the treatment; when the fact is, that is might be faulty design. I don't think I would throw out the results just because there wasn't enough power to the design.

Feldhusen: I do want to come back, then, to what I think is another look at the problem itself, fundamentally. I question strongly the assertion that it is built into the nature of things—fixed immutably—that you can only work with small groups, so that you are forever plagued with this problem. Quite to the contrary, I made almost nasty assertions there at our morning break to Rosa, about why I think this comes about in much, I think, special education research—such things as people fortuitously, or conveniently, coming upon a small group about whom they happen to have data and can do an easy follow-up. That sort of thing may still provide valuable and useful insights, but it does not constitute carefully planned long range research. It is a nice thing to capitalize upon an available small group. Surely, with aggressive, disruptive children, there must be many such studies I can document, where N is 10, or even the studies reported in the Journal of Applied Behavioral Analysis, where the N is quite frequently 1. But, in any event, I just can't believe that you are doomed forever to small Ns, if people set out from the beginning to plan good, large scale research, that they can have decent Ns, and not be plagued by this problem. Is this so, or not? Now, you may react.

Mann: Well, the fact of the thing is, in special education, most samples are small.

Feldhusen: Do you agree with my assertion this is, in part, due to people simply not being ambitious in getting an adequate number of subjects?

Mann: I think, in part, it is. But in special education, it is also genuinely harder and more expensive to establish a large N than in other areas of education, and you can't always find the bodies.

Millman: It will be tough in any field, if you want to say I want to do a research study which only includes 13-year-old boys who pre this and post that and have scores of 68.7 on a 7 point scale. You know, you can define it so narrowly that, of course, you can't find a body and I guess my answer to that would be that it—to repeat my faith—is that you can find results you are interested in that will probably cut across a much broader range than just this narrow slice. I can't conceive a treatment that will work with this kind of specific person and not work a little bit on this side and a little bit on that side of the original definition of subjects, so that you can get more bodies. Too narrow a definition of subjects seems to have caused small Ns in many cases, and unnecessarily.

Mann: Well, again, the research in the field is directed to certain categorical groups. If you have dyslexics and you want to study dyslexics, to throw in an EMR who also can't read distorts our group composition. This is why we often use such restrictive definitions of sampling. Your point is well taken, Jay, however.

Millman: I can see a resistance to mixing together, for example, steak and pears, but I'm not sure I see the resistance to putting together pears and apples.

French: I understand what you are saying. I think that there isn't anything that money couldn't overcome. It would involve, in many instances, cooperation across school district lines, which is a tremendously difficult political thing to do. It involves the researcher putting a priority on doing that kind of data gathering and adjusting his time to be able to do it. I suppose that if there was enough money to employ researchers to do it, it would get done, but the priority of things just has not been that way.

Millman: I think before I resigned myself to small Ns as the inherent nature of special education researches, I would want to take a hard look at the learning path, and say, really, why do I think that this is only going to be effective with dyslexics and not with some other type that might be available that could push up your N.

Mann: Rosa? You could speak on this as well as anybody.

Hagin: Well, I think you have to first talk about the kind of child you want to work with. If you pick something that is a rare defect, you are forced into small Ns. If you are going to work with phenylketonuria, for instance, or if you assemble 30 Turner Syndromes—this is a good N for Turners—if it is going to be children with language disabilities, you might expect a better sample. And so, I have the same problem as you are troubled with—the conclusions with small Ns. I am also troubled with the very large samples that mask the true nature of the problem.

Mann: I think you have expressed Bart's question, and also mine, more effectively here—that you have a large N that is masking all sorts of deviant groups within it. Now, you are saying this isn't really critical to the research, Jay? If you want to find out how MRs are...

Hopkins: ...meaningful subpopulations? Well, then you categorize them. You know, factorial design strategically...

Proger: If you continue sub-categorizing, your cell sizes become ridiculously small.

Hopkins: The size of the N per cell is an irrelevant consideration, in terms of power.

Millman: I think that point needs to be amplified.

Hopkins: Intuition leads one to the position that the number of cases per cell is a critical building block in the adequacy of the study, but just because, let's say you have 100 students: you categorize them by sex, and then their full IQ levels, and the two ethnic groups, or something. That doesn't mean that the quality of your study has suffered because now you have 5 per cell instead of 50. In fact, you probably have a much more efficient study, in terms of the overall main effects, plus you are able to see whether or not there is interaction between treatment and all of these other factors, and if there isn't, then that's your statistical basis for generalizability. If there is, you found out that it is worthwhile for everyone, except the CP group.

Mann: You are saying the N of the cell isn't important?...too important?

Hopkins: In a comparative sense, you've got this group of kids and whether you group them into small units hasn't decreased the quality of your studies; it has increased the quality, both in terms of power and generalizability.

Feldhusen: May I probe that just a bit with you, to be sure I understand, as I think I understand? Let's separate ourselves into tests of main events and tests of interactions. You said something between there that I missed. Let me make an assertion and you agree or disagree. You are minimizing size in the test of an interaction of a particular subgroup with the treatment. However, you would prefer that the N in the cells involving that subgrouping to be larger or to have a stronger faith in the reliability of the interaction. No?

Hopkins: All things equal, sure...as large an N per cell as possible.

Feldhusen: I just...okay. I want to be sure that I don't hear you saying that the size of the cell is irrelevant.

Hopkins: Given this set of data, do we lose information, or have a less efficient analysis because we categorize them into more homogeneous subgroups by age, sex, IQ level, or whatever, and so we only have two observations per cell, where if we didn't categorize them we might have 100 in the two groups? Of course not!

Feldhusen: You are saying we do have a more efficient analysis, now.

Proger: ...given that number of bodies.

French: Yes, but there must be a reason for putting the subjects in the cells—no sense in dividing them into boys and girls, if that isn't relevant to the task.

Hopkins: Yes, but you may not know that a priori. My original comment was, if they are meaningful subpopulations that we are worried about, in generalizing the total differences of ways to all kinds within that group, let's categorize it and see whether or not...

Proger: You will always extract more error variance the more you build into it that way.

Hopkins: It may. There are other...

Proger: But again, in terms of generalizing, if you get an EMR population, you may have sampled from it unrepresentative members of the group as a whole, because IQs for one range—from, let's say, 50 IQs to 80 points—30 IQ points spread is considerable. And the cell you have established may be very different, in many characteristics, from the population that is important to us as special educators. So, your suggestion still leaves us with our problem.

Millman: Maybe one way to make Bart's point intuitively clearer is to say that the reason people may be concerned about the existance of these specific meaningful subpopulations is because the effects that you find in your studies might not generalize to all of these kids. It might be okay for one kind of EMR, but not for those EMRs who are not within the narrow range represented by the cell. Another way to phrase Ben's approach is to say that maybe there is an interaction that is existing and you can say, okay now, how do we measure whether there is that interaction? And if you look, the test of the interaction has approximately the same number of degrees of freedom for testing it as the total samples that you use, not for the number in the cells. In other words, the test is based, not on the number of kids that you have a particular cell combination of, but is more related to the total number of kids that you had before you divided them into these units of population. For that reason, you don't give up too much by making that kind of categorization. I don't know if that helps.

Hopkins: It does. Indeed, you may not give up anything if the variable in which you are categorizing is correlated with criterion performance. In other words, if there is a sex difference, then you employ sex as a factor, and you may gain more power than you lose by breaking them down into...

Millman: And then, if you have as many as 20 kids or 80 kids, then, of course, the difference in degrees of freedom is so slight that, you know, you are sure by doing it. I think a good operating rule is, almost always, whenever you can divide your example into a meaningful subpopulation.

Feldhusen: In preference, also, to the other suggestion made here earlier that one could take the covariance route here. If one could get an assessment along some continuum that would apply to the whole, then one might be tempted to remove it, or to bring it under control by a covariance design. I think I would infer, from the exchange here, that identifying the subgroups and carrying out the kind of an analysis Ben has been talking about here, would be preferable to the situation where we could use a covariance design. Is that?...I believe that is so, and I would be happy to go on to justify it, but I am happy to hear your ideas.

Millman: Well, I think—again, checking on this—studies have shown that when the conditions for covariance have been met, that you are better off doing the blocking provided thatthe relationship between the variable that you are using as a covariant and the dependent variable is not up high, say above 47. It depends also, on the number of treatments, but I think if the relationship isn't really high, you are probably better off blocking. I'd say even if it were very high, you would probably be better off blocking than using covariation because the interaction would slow...in covariance. Ken, you're...

Hopkins: I agree completely. There are two issues. One is statistical efficiency, or power, and you can get more power by covarying if there is a very high correlation between the covariable and the dependent variable. On the other hand, what you lose is the ability to detect interaction, which is really the heart of the issue we are raising. It isn't just power, it is a matter of generalizability of the finding. Can you generalize across the whole ability span that you sampled? So, I agree completely with Jay, that when you are in a true external context in the Campbell-Stanley sense, it is certainly best to stratify, or block, or whatever you want to call it. On the other hand, if you are in a quan-experimental situation where you have to compensate for bias, that throws a whole new dimension, and you would lean toward the covariance, I think, in that situation.

Bracht: I would like to bring up another issue. There is, I think, some suggestion to have multiple outcome measures, rather than one outcome measure. How do we handle that sort of situation? And also, how do we deal with some of the things mentioned this morning about observing what occurs in the treatment if something is then more effectively in one treatment than in another treatment? How do we treat this type of thing in the analysis? Or, if the teacher is more effective in one treatment than in the other, how do we handle that? Or, if students are more disruptive in one than in the other? These are very important things in the study, but they seem to be ignored when we do our statistical analysis in most things, and yet, they may be the critical factors that have brought about the results. How do we handle such things?

Millman: It sounds almost like the problem is in defining the experimental unit. You seem to be saying that one of the problems may be that although we got a difference, it might be for any number of reasons: that such a destructive kid, or a particularly talented teacher, or something, rather than what we want to consider as the cause of the effect, and the problem, of course, is that these things can be the determining factor. Don't rely just on one teacher teaching it; have different groups of students, not just one.

Hopkins: I used to kind of take the position that in almost every possible study you'll get information, even if of low quality, that is of some benefit to you. I am now kind of changing to a different point of view, because of the noise that is injected by inefficiently carried out studies. I think if you get a treatment where the teacher can be a critical variable and you only have one teacher, you don't have a study in my opinion. If you can't have more than one treatment, or more than one teacher making application of the treatment so that you can see whether or not there are significant differences as limited as that is, to those two teachers, then I don't see a study.

Millman: I'd go a step further, to be of help to people who are in this situation. I would suggest that you really think hard about the extent to which you think some things are really going to make a difference. Since it isn't that important whether you've got 8 kids together or just 4 kids together for a cell, I'd say instead of one teacher teaching the 8 kids in this particular way, divide them up into maybe 3 groups of 3, or 2 groups of 4, and try to vary the conditions that are not the crucial treatment conditions. If you don't I think you're in a very difficult position of having so much confounded in there, that it is going to be hard to do very much.

Mann: Now, the chances are—it does happen in randomizing groups and whatever—that you have one cell where, lo and behold, the kids really are pretty bad. You may have randomized in this and that, and you have identified various groups, but one cell is miserable; it is a real miserable cell. Now, what do you do? Do you just dissolve the whole experiment and do it over again? Do you rationalize at the end? Do you adjust the probabilities, or what? Do you replace the cell with similar Ss waiting in the sidelines?

Hopkins: Well, how do you know it is miserable?

Mann: There are, in fact, events beyond control—that we just send one group away. I've seen it happen.

Hopkins: You can still covary on it.

Mann: You think that is the answer then, for this sort of thing?

Hopkins: Well, the answer is to have potent treatments.

Mann: But, let's say pretest data doesn't specify, say, aggressive behavior; you have EMRs, so you divide them up and you find one very aggressive EMR group not responding to treatment because of this, and one that is very docile and respondent to treatment B, and so on. You can't really covary on this, I wouldn't think. The pretest data didn't take this into account.

Millman: First of all, if you can anticipate about what is going to really screw up the treatment, you know, like agressiveness or something, you try to get information on that and make use of it when the time comes. But suppose that you are not that smart—few of us are—and this comes to light after the thing gets started. Well, if you are willing to make the assumption that the treatment did not cause the aggressive behavior, then what you need to do is to get normal 1-0—is or is not—on each kid. Is he, or is he not, a particularly aggressive kid? Something like that. Even if that is all you can get, you can make use of that.

Hopkins: Can it be treated as a classificatory effect?

Millman: Well, that may be hard to do.

Hopkins: If they are all one group. I think my own personal feeling about most examples like that is that they are artificial. That is, I think kids differ on a continuum on aggressiveness and if the worse kid is over here, then there is someone in the other group that isn't that far off from him, unless, of course, you have very small Ns, in which case you are not going to find too many differences anyway, unless they are overwhelming.

Millman: Take two aspirins. We'll be back.

Feldhusen: Let me ask a question here. Suppose we have five children in a particular classroom and things go beserk there, somehow or other. And this is related to one of your treatments; they are all within a particular treatment condition in that classroom and things have just gone to pieces completely. Now, how would you react to doing the experiment over, for that portion of the experiment? If we have a population from which we can randomly draw yet another of five, and can randomly find another classroom to put five into, how would you react to such patching up? Let me say that the condition bears no relationship to the treatment; it came about suddenly, just as our experiment started; we did not know it but the school system brought in a new teacher.

Mann: Or, the kids were taken off drugs that darn day. The physician took them off drugs. They are acting out badly. It can happen, and it does happen.

Feldhusen: And furthermore, we want to generalize to other classrooms and kids, where teachers are using my material, my methods, or what have you. And now, as I said, the teacher suddenly—just at the start of my experiment—was yanked out and I didn't know about it until after I cam back six weeks later. Now, are you content to patch up, the way I suggest, by trying to replicate that piece?

Hopkins: I think what you are suggesting makes a lot of sense, but I would like to make another observation, to kind of put it in context. I think that, in education, we are much too impatient and after instant solutions that are going to be yielded by one study, I think we could take a lesson from psychologists who may have a series of research studies in a given area, and the imperfections they learn in one study they prevent in the next study, and have a sequence of studies on a given area.

~~Proger:~~ MANN Let's say our resources don't allow you infinite series of replications. Research in psychology is greatly time-demanding than that in applied education.

Hopkins: Well, it isn't just a replication. It's found that this other variable that you did not anticipate—aggressiveness, you know, is a problem—so next time you can get some measure of that. Teachers have identified them in terms of past behavior and built that in.

Mann: John has given us a very good example, over here. Now, is this a legitimate thing to do? We have the study; it is costing a bundle; this is a one-shot affair. One cell went beserk—no control. It is not our fault. What can we do? Can we patch it up as John says?

Feldhusen: Instead of saying legitimate, say it is even barely tolerable.

Millman: Yes...I would say it sounds terrific and if we could do that, by all means go ahead and do it. Obviously, in the reporting of it, you would need to indicate what you did and any reason for believing that this group initially was different from the other groups.

Feldhusen: You would see immediately, in relation to this, the two internal validity hazards of maturation and history. One, the treatment is now going to go on at a different time and the original population—if you sample out of it—has grown older a bit. So, both of those might be minor hazards.

Mann: Would you accept that, Ken?

Hopkins: Of course, all of us do the very best that we can when we are in situations where things don't go according to plan, and that sounds like about the best that you can do in that situation. On the other hand, don't ask me to say that what you did really satisfactorily answers the question you started out to answer. You may have more confidence in the answer that you have now, than you would have without that study, but more work is needed. I think one would become very frustrated with the whole research strategy if he is expecting it to have a very high payoff in a very short interval.

Mann: Let me ask another naive question along the same lines. Let's say you are just doing an experimental study—small cells. One very disruptive subject again appears within one cell—just one. Would you allow the researcher in special education to pull this one subject and and dispose of him, by saying he really didn't meet the criteria, etc? It is done, of course.

Hopkins: Well, first of all, you can do whatever you want to. If the person doesn't really belong in a population, sometimes you can kind of justify that statistically. There are tests for outliers and tossing out extreme cases and controlling the range, and it gets very messy, but what should guide us is getting to what we think the problems are, and next time preventing that kind of person getting into our study.

## REFERENCES

Campbell, D. R., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.) Handbook of research on teaching. Chicago: Rand McNally, 1963, pp. 171-246.

Clasen, R. E. Book reviews: An introduction to educational research, third edition, by Robert M. W. Travers. Journal of Educational Research, 1972, 66, 56 + 60.

Gage, N. L. Paradigms for research on teaching. In N. L. Gage (Ed.) Handbook of research on teaching. Chicago: Rand McNally, 1963, pp. 94-141.

Gottman, J. M. N-of-one and N-of-two research in psychotherapy. Psychological Bulletin, 1973, 80, 93-105.

Poulton, E. C. Unwanted range effects from using within-subject experimental designs. Psychological Bulletin, 1973, 80, 113-121.

## CHAPTER 16

## MATCHING AND THE CHOICE OF APPROPRIATE CONTROL GROUPS

Overview: In the history of behavioral science, the selection of appropriate standards of comparison for the performance of an experimental treatment group has seemed rather straightforward. There are, however, several fine points to be considered. In a "controlled" experiment, one is interested in being able to attribute the apparent effectiveness of a treatment to the treatment procedures themselves and not to various sources of error. Thus, the control group of a no-treatment or alternative-treatment nature.

One of the original methods of building treatment and control groups was to use classical "matching" whereby one included subjects in each of his groups so that they were apparently very comparable on one or more criteria ("eyeballing the data"). A major problem here is that the precision or closeness of the matched groups is usually not specified. Further, one should not match on variables (such as IQ) when these are available to more directly related measures of the skills in question. In more recent years, investigators have taken to stratified randomization to yield a randomized blocks design and meet some of the objections of traditional grade matching.

Some general points about forming control groups are discussed. First, it makes little sense to have no treatment control groups. The control groups should be given some alternative, structured program. Second, if one has designed a treatment for use with a specifically handicapped population, we should not have control groups of widely differing characteristics. Third, in more applied research, such as program evaluation where one often does not have the freedom to form nicely structured control groups; the retrospective controls can be established on the basis of the previous year's group of children who may have used a different program of instruction. Fourth, the Hawthorne Effect is discussed and realistic ways of grappling with it are given. Fifth, if one looks upon educational research as a progression of planned, systematic studies, then one need not design, as his first test of a specific treatment, a single grand multiply-controlled study to assess the input of all variables. In the sense of formative program evaluation, one should be interested only in obtaining an overall treatment. Later studies can isolate the sources of the significant differences. Finally, the ethics of experimentation arises in connection with control groups. Is it ethical to deny control subjects an apparently beneficial treatment? For treatments that may have dangerous side effects (such as drugs), is it ethical to subject experimental groups to the procedures?

**Mann:** I'd like to discuss the old issue of matching again. It is still done, by and large, in this field. So I think, somewhere a statement concerning the perils and the fallibility of it is certainly valuable.

**Feldhusen:** May I ask a question here? It does seem to me that there is, or persists in the literature, a use of the word "matching" which is an acceptable procedure. I believe it is Edwards, at least up to the edition of the mid-60's, that uses the word "matching" but, I believe, in a perfectly acceptable way. Lester, you think you are talking about trying to eyeball one's sample to find somebody over here who seems to look like somebody over there, and placing one person in an experimental group and the other in a control group. That is called "matching." But other uses of the term can be found. I believe Edwards persists in using the term "matching" to refer to a condition where I would have a single score—or a composite of several scores—for each person in my sample. In either case, I have a group of people whom I can arrange...this is my total sample now—undifferentiated for experiment and of control...downward from the highest to the lowest score. I can now randomly proceed down through this group of people and assign one person to experimental, the next to control, back and forth, etc. I believe Edwards used the term "matching" for that procedure. The latter thing I have described, I believe, is an acceptable procedure to forming experimental groups even though it might be called "matching."

**Hopkins:** I agree with John that a lot of people refer to that as "matching," but I think that confuses the issue. You know technically, the randomized Blocks Design is a form of matching. There are defined groups of population and you try going in and find comparable individuals.

**Millman:** The principle distinction that I think must be made is whether the treatment is—whether you assigned your treatment after the matching or before the matching. In the first sense, you had an intact group, that is, groups already formed—one destined to be the experimental and the other predestined to be the control. That is quite different.

**Feldhusen:** Lester's concern, I believe, is about people matching on the basis of their eyeballing data. I did that myself years ago, and it's horrible when you are speaking "matching" for, let's say, IQ Match-trys to find somebody over there who looks pretty close to someone over here and then assigns one to the experimental and another to the control group.

**Millman:** How is that assignment usually done?

**Mann:** Well, if you compare, let's say, normals with retarded, you might match on IQ and let the CA run.

Feldhusen: This is the case of clinically distinct groups matched precisely on one variable.

Mann: One variable...but you can match on more than one variable. You want to establish control group A and experimental group B. You get the girls from Mrs. Sue's EMR class and get one master list with columns of data on IQ, CA and sex, let's say, and then you match this intact group of EMR girls with a similar list from Mrs. Sally's intact EMR class.

Cross: Let me change the topic slightly from the specific issue of matching to your broader issue of what an appropriate control group is. Do you think it is appropriate to match children with different chronological ages but having the similar mental age, which you see over and over? It is standard practice but I reacted to it negatively. I just didn't think it was appropriate but I didn't see an alternative way out of it. It just seems like that is one of the real hang-ups in this matching.

Mann: That is another issue. A lot of the work that has critically influenced the field is based on this sort of thing.

Cross: I don't know a better way out of it.

Feldhusen: Would you restate that again?

Cross: Well, the typical control group for a study with EMR's, let's say, is a normal group with the same mental age, but often much younger chronological age and, of course, the MA which they are seemingly matching represent different processes in the two groups.

Feldhusen: Could you state for us the type of question the person is trying to answer in the situation where this methodology is being applied? I am hard pressed to understand what kind of a question the researcher is trying to answer here.

Proger: It is certainly a crucial issue. I am glad you raised it. It comes under the old control group—appropriateness problem—but it is not really a matching issue, as such. I think the basic purpose for such a study would be developmental in nature. You want to see if the educably mentally retarded do thus and thus in, let's say, treatment A vs. treatment B as an issue unto itself, but also compare those findings with what would normal expectancies be with a similar mental age but a lower chronological age. If they are normal kids, the only way you can get a similar mental developmental level to the EMRs would be if they are much younger chronologically. You could also have another normal control group in the other direction with the same CA but much higher MP. It is matching, in the sense of the multiple-criterion problem raised by John. That is, you are varying to match on CA, MA, and maybe sex.

Miliman: One could certainly say, in terms of matching, that having a control group is better than not having a control group. It gives you more information of comparison. Now, there are many ways that one can form control groups. Some are better than others but having a control group which is, in some sense, like the experimental groups in a variable not directly being manipulated or available, if it is part of the experimental treatment is a desirable thing. I think one could say that. Now, there are many different ways of forming such a control group. Some ways are more satisfactory than other ways and other ways are still better, even though they are unsatisfactory than perhaps not having any control group at all. The most satisfactory way of forming a control group is, in a two-group study, for example, to try to make the two groups as equivalent as possible. You match on those variables that are apt to be related to success of the experimental treatment pairing anyway you can to make them as equivalent as possible. Then, if you will, assign randomly which member of each pair is to get the experimental treatment and which does not. If you are not able to do that, that is, if the individuals who are going to be in the experimental group are already determined and you just can't randomly assign who will and who will not get the treatment, then you still go ahead and try to match if there are some problems there. Ken wrote, I thought, an excellent paper published in the Journal of Special Education.

Bracht: I was just wondering about the problem we were discussing where we were identifying possible various control groups: one in similar chronological age but different mental age and then similar mental age but different chronological age. Are those groups obtained by matching individual for individual or just defining characteristics for a control group? For example, MA must be below some level or within a certain range.

Mann: Both are done. A very common practice is, let's say, you have these EMR youngsters and you go to the normal population and you choose children who match up on sex and IQ. They are selected on that basis. You are looking for people who match each other in pairs, triplets and so on. There are other cases, obviously, where you have this one group of a certain type as compared to another group without any attempt at individual matches, but the group measures are one variable or another one comparable.

Proger: I think, though, I might not agree completely with what Jay said about the idea that an appropriate control group in an experimental study would hopefully always be a comparable group of people to the experimental subjects themselves. I think the point we were trying to make here, when Larnie brought up the other type of control group, is that this is a totally different type of study. I would envision that you could have Jay's situation but the second type is a much bigger elaboration of his. It is saying that we will have a comparable group of kids under Treatment A and maybe Control and Treatment B. That will hold under his criterion of a control group, but then we will go further and actually try to get an experimental definition of retardation. What happens when we compare those groups with the normal MA counterparts but lower CA and then vice versa with the CA but different MA? I think that is a bigger design; you would have the classical comparability within a level of retardation or normality but you twist it more ways. I think, again, it is for developmental psychology purposes, which is interesting in EMR research. I think it is a control, too, in a grosser sense in that it does lead to the self-evident hypothesis in that you would expect hopefully the normal comparable MA's to do better than the similar MA retarded, perhaps. You would hope some things would come out this way.

Mann: Let me give you an example that just came to mind, in terms of the MA matching of a retarded and a normal group, in studying mediational processes. The researcher wants to see whether this is a developmental issue, really, or is it simply a matter of state and time and experience. You match up for MA then; presumably the retarded with an MA of 9 will have similar problems, in terms of mediating, as will a normal average 4 year old if the MA's are the same. That is a typical child experimental paradigm method in the field, wouldn't you say?

Feldhusen: We seem to have accepted going into two distinct populations here...let us say into educable mentally retarded and into a normal population and from each, trying to find peers or individuals with the same MA.

Hagin: I would like to know what this MA represents.

Mann: It comes from standardized tests and this is how they do the matching—either the Binet or the WISC, if it is an MR population, as a rule.

Hagin: Yes, but you would realize that David Wechsler has a very strong case not to use the MA. He feels, and you can read this in his manual, most philosophically, that this is a strawman and it means something very different at different age levels.

French: He did put a table of age equivalents in the back.

Hagin: Only after user demand!

Mann: In terms of MR research, however, they have found the MA useful. I think that they use MA and IQ as presumably different things. MA may represent learning and IQ the rate of learning. This has been suggested.

Hagin: Wechsler has called the MA a "bastard" measure, being of uncertain parentage and I think we use it inappropriately. It is just a test score. It is not an absolute level.

French: He also said at convention...I guess it is not a secret... I was sitting among several people from Psychological Corporation and they were all cringing. This is almost a direct quote: "The Binet is a good test. I can afford to say that now." And he admitted it could be used for MA purposes. Instead of using an age equivalent, he would say to use the raw score from each subtest and use that for the intellectually normal kids, as well, in forming groups such as this. This is begging the question, because he has also said that the Binet MA is also the raw score times two.

Feldhusen: But we are gathered here to talk about research which is supposed to find empirical answers to these kinds of assertions that Mr. Wechsler and others make. I think that it is perfectly reasonable to conceptualize questions about MA. Do, in fact, normals and EMR's who have the same MA's perform equally well on the same tasks across the tests?

Mann: On certain tests, they do.

Feldhusen: That, of course, is established; that, I know. So that seems to be the nature of much of this line of investigation, isn't it?

Mann: Much of this is done upon this kind of matching. I would say on establishing "equivalent" different groups.

Feldhusen: The purpose, then, is to determine whether it is not actually different behavior—different performances—which lead an EMR to achieve this MA level as opposed to a normal who achieves the same MA level but is performing more efficiently, or less so, on particular tasks.

Mann: ...but also to compare normals to retarded. For a long while, in physical education research, there was the delusion that retardation wasn't related with intelligence, except in extremely low groups. So then, they started to match up these youngsters, EMR and normal, to see if, in fact, they are different and, lo and behold, they were. So, you may have matching to study processes and also matching to establish groups upon which to attempt experimental treatments.

French: I think that one of the questions that will be asked, on a continuing basis through the next few years, is what are the characteristics necessary to learn a particular task? A number of people say we shouldn't classify children as mentally retarded or emotionally disturbed. We should just describe their characteristics and plot learning programs for them and it seems that it is a very reasonable question—whether we should still continue to use MA's, IQ's, or other scores from a Wechsler, a Binet, or some other test, or to get off these into other characteristics of a learner that may be equal.

Mann: Can we establish some statements or resolutions from the people here, in terms of, let's say, the general issues of matching for whatever reasons, now that we have a number of bases upon which to match.

Hopkins: I have some general reactions to the notion of matching. That is, using the term in the sense where you have a defined population and you are going in to try to remove an extraneous source of variation from your results, let's say, by matching mental age. I agree with Jay that usually any kind of control group is better than none at all, providing that the user isn't beguiled by his use of the control group. I think that one of the unfortunate things about matching is there are sources of contamination that are not apparent intuitively. They come in and cloud the picture and beguile the investigator and confuse the profession and that presents a problem, not of using a control group, but a misinterpretation of information. The manner this comes about consists of more than one way. First of all, most matching studies I have seen never indicate how closely they come when they match on IQ. You don't know whether it is matched within 2 points, 0 points, within 10 points, whatever. That means that, in many instances, there are significant differences on the matching variable between the groups. In other words, when there is a large N, will there be a significant difference on the matching variable before the study is ever conducted if the usual matching operates as coming within 5 IQ points? Such matching, to begin with, hasn't eliminated the source of variation. The solution is not, then, to match on identical scores: identical MA, identical IQ. Of course, that would give you zero differences between the groups on observed scores. But that is not really what you are interested in, either. You are interested in matching on the ability in question—not general potential—and when you match on observed scores, you've got measurement error involved in matching. Usually, the reason you are trying to match groups is that they differ in central tendency on this measure and so you are matching brighter than average delinquents, most of the time, with duller than average non-delinquents. And if you retested those, they would regress towards the mean of the population they represent, and without treatment, you retest—you get a significant difference between groups on matching variables even if you matched on identical IQ scores. So, again, it all depends on the user. One helpful suggestion is what

McNemar suggests, that is, to match on predicted true scores. You would never match on observed scores. That wouldn't be as good as matching on predicted true scores which is more troublesome reliability. Only group, if you don't have a good estimate of it and then use that to predict true scores.

Mann: Would this problem exist if you were matching from within the same population, let's say, just within bright delinquents?

Hopkins: I think it comes back to this whole question of how we are using the term "matching," as John pointed out earlier. Apart from classical matching, you have the randomized design, as John was saying earlier, where you order them on something like mental age and then you assign two of the top pair to experimental—I mean the top quartet—and the other two to control. Actually, there are no problems at all there. That is the ideal strategy, when you can implement a blocking situation. But, if you are comparing blind with seeing youngsters and see if they get the concept of conservation at the same time, and see whether or not visual stimulation is necessary before they get conservation of matter, you have no alternative but to try to compare one group with another group—the designation being not amenable to manipulation—and there matching is more difficult.

Millman: Well, I think Ken has summarized some hazards of the question. I would like to give my opinion, I guess, about the question of whether you should compare, say, the mentally retarded with a normal group. My feeling of the matter is that the answer is "yes and no." If you are interested in the effectiveness of some sort of intervention program on a, say, group of mentally retarded youngsters, then I am not sure how worthwhile it would be to have as the comparison group, a normal population. I would say, probably in that situation, there is patently little to be gained. In a developmental study, in which you want to answer some of the other kinds of questions that have been brought up, perhaps there is some more justification for comparing them with a normal group. I would just say that I think the value of developmental studies has been overemphasized and that knowing what kids can do, or what kids do at a certain age, isn't too exciting. I also want to raise a quite different issue. You know, when you find that mentally retarded children possess this kind of skill, or think this way about this or that, and normals who are the same mental age react this way, that doesn't mean that it has to be that way. It might be an interesting fact, but that doesn't say, that with proper intervention it must remain so. For instance, if you find that blind people don't have the conservation of energy when compared with non-blind people of the same mental age, that doesn't mean they can't have it. It just means that given their educational environmental history, they don't have it.

That may be interesting to know, but I don't think we should then say, "Well, don't bother to teach conservation of energy to blind people until the MA of 7-2, or something." That isn't, it seems to me, a justifiable kind of practice to take, on the basis of a developmental study. While I think comparisons are okay to make, we must be careful not to take what is, as what must be.

Bracht: It is important to make the distinctions of "experimental" and "developmental" that Jay has made. I think we are asking different questions. I think, that in the developmental studies where there appears to be some need to match subjects from different populations, that this is the most useful approach to take. It is better to clearly define the populations and do studies within those populations. You can still make comparisons. It is not truly, in all cases, a formal comparison but it is useful information from which we can make some logical analysis.

Feldhusen: This is a very small matter, but yet, I think very important. I think we have, through our conversation, established that it will not do for you, as journal editors, let's say, to accept a manuscript which says we matched on such and such a variable. I think that it should be a rule that more detail is needed so that we may understand what actually took place because of several different things, some of which are troublesome and some of which, as Ken's established, might be perfectly right and appropriate. Good procedure might be subsumed under this thing called "matching." So, we need more detail than the simple assertion that we matched on IQ.

Hagin: There was something Jay started and then we got talking about something else. He said that it was the manner of assignment to experimental control group. Can you elaborate?

Millman: Well, this goes back to what Ken followed up on. Ken said that if the manner of assignment is such that you, in a sense, form your matched pairs, say, and then randomly assign the treatment—if that's the manner in which you do it—there's no problem. I think, to use Ken's words, that's the ideal situation. Whereas, if your manner of matching is to deal with two populations that are already formed and then try to match in that way, then there are all kinds of hazards of which you must be aware. It's still good to try to match, but don't be beguiled. That's all I meant.

Feldhusen: It is really becoming much more, I think, well established now to refer to this good form of matching, as Ken has asserted, as randomized block design, and not to persist in calling that matching.

Mann: John, what are the ethical problems involved in denying control subjects an apparently beneficial treatment?

Feldhusen: I don't know what to say about that. We all know that there is the problem of not just blithely withholding what might be an efficacious procedure treatment from a group of people, and assuming we never have any further obligation to them, if they've been identified and have served us as a control group on whom we've made assessments. I really don't, at the moment, have anything further to say. I guess we have an obligation, if at all possible, to try and provide the treatment later to them if we possibly can. Now, I don't believe that our committee at Purdue University concerning the protection of human subjects in research—and I guess that descends from the NIH dictum —presses us to do this and I think it would oftentimes be right. But, we do not have the resources; we cannot offer and are not being pressed to offer the treatments of intervention at some later date and, as a matter of fact, in some cases, it would become meaningless. In our studies of tutoring, the people go into a course and the course transpires. We operate over a semester with our intervention tutorial system and, in a sense, it's too late to do anything for them when the semester is over. Again, our whole procedure has been submitted to the Committee on the Protection of Human Subjects and no one has objected to the fact that we don't intend to help those people who are the control group in any way. And, I could see, oftentimes, that one couldn't. We would not have the money; it costs us quite a bit to provide the treatment; we would not have the resources to do it for the control group.

Mann: Joe?

French: Well, I don't know. I was mulling over some of the ethical issues involved there and I keep thinking of just how important it is to have a control group to verify the effectiveness of the treatment. I'm thinking of the times when the ASNIK vaccine was used with school children. Half of them got the placebo and half of them got the real thing. I think it's too late to go back for some of them. We have to have a control group to determine the efficacy of various things. I think the only thing we can do is make sure, when we're concerned with issues as grave as treating polio, that we do everything we can to make up the difference to the control group. But, for many of the studies that we conduct, it's been a guess as to whether or not what we do for the experimentals is going to be helpful to the children. I guess there's a rule of practical significance. We may prove with the ± test that it's significant, but it may not be worth going back and trying the same procedures with control children. I'd hate to be bound by such an across-the-board rule as trying to provide treatment late to the control group.

Mann: Let me put another question in here. In special education, we have particular problems because of the nature of the population that we work with—the expectancies and hopes and the magical qualities of our interventions for many people. With this in mind, is there really a control group, or is it a fugitive treatment group, so to speak?

French: None of my thoughts are in the direction that you're trying to take it. I think we have to be very careful about the control group, in working with handicapped kids, because we don't know what the parents are doing for the kids when they're not in school. It could be we assume they're a control group and they're not getting training and activity that's related to the experimental group. It may very well be that the parents are securing a tutorial program, or have something after school, that would violate the effectiveness of the control group.

Feldhusen: Well, the opposite site of that coin—which I almost thought I heard you implying—was that powerful expectation might emerge from a particular treatment with a special education group. We worry about something like a Hawthorne Effect in other research, so it might also be an extremely powerful thing, operating in special education. If we talk now about a control group that remains essentially ignorant of anything that is going on and, let's say, that they don't guess anything is up from the pre- and the post-test. Of course, what IQ they receive won't make any difference. However, if such is not the case, then sorting out treatment from this Hawthorne Effect certainly seems to suggest calling for an ACE treatment in between. Now, this really gets troublesome.

Mann: No double blind is possible here, of course.

Feldhusen: Yes, how awful—really it is—to use a placebo treatment that would induce equally powerful expectations of cure, when indeed, you knew if the research came out right everyone should be doomed in that particular control group to failure. So that also is a slant on that third type of control group that is really ethically troublesome.

Bracht: I think control groups are more important in some types of study than in other studies. We were, yesterday, discussing that some of the research is related to a complex training program. I think control groups are, in many cases, not highly useful in the initial stages of the evaluation of some programs. Quite often, when we're developing a program, there are a lot of bugs that have to be ironed out initially. Quite often, a new program is not highly effective and does not reach it's maximum effectiveness in its first tryout. I think, at that first stage, a control group is not as useful as, perhaps, at a later stage. We first try out a program, do a thorough formative evaluation, and then maybe compare the program to a control program. I think, here, we get more usefulness from the control group. Another comment about control group is: What is the nature of the control group? Is it a group that receives no treatment, which, in one sense, is what a control group defines? Another: Is that control group a group that receives some alternative treatment from a newly developed program? I am, personally, more interested in how much we can improve over what we're presently doing, and would prefer the control group to be, in most cases, what we're presently doing, and can we improve upon that with some new program. I think, to have a control group as a group that receives nothing, sometimes provides very little useful information. We might have a study here where: treatment A is better than nothing received by a control group; another person finds that treatment B is better than a control group; another person finds that treatment C is better than a control group. But we have no idea, yet, how these treatments: A, B, and C, compare to each other. I think that we might provide more useful information by considering control groups as some current programs.

Hopkins: Well, just a little bit of a comment on the other side of the fence. I think, often, one assumes a priori that the treatment is going to be valuable. There is not much of an ethical problem here. However, in Joe's example of the vaccine, you've got to ask, Who's taking the risk? Just recalling the early smallpox innoculation, it was a treated group that took the brunt of the bad outcome. So one can say, on the one hand, is it ethical to use kids as guinea pigs? That's exactly opposite to the position: Is it ethical to withhold treatment from kids? I think if you know it's efficacious beforehand, then there's no point in the study. But if you don't, you're not sure who is taking the risk. So, I think, that kind of balances out the picture. Although, as someone mentioned yesterday, one way to solve this problem, in certain situations, is what Don Campbell told us 2 or 3 years ago: to try to elicit more interest in subjects among the population than you're able to serve. Then it seems quite justifiable that some people must, by default, go into a control group.

French: I'm disappointed every once in a while, in reading articles, to know that there was a Hawthorne group but not to know what they were doing. I've heard of some situations—it was for a visual perceptual program of some kind—and the Hawthorne group was playing Old Maid. I wonder about how different the Old Maid cards are from the stuff in the training program and if that's an appropriate Hawthorne Effect. I don't know how you check that, but, I guess, we need to know more about the Hawthorne groups that we often find in the literature.

Feldhusen: I'm going to go back and say I am reminded of Desmond Cook's assertion that if this is really such a big thing, why don't we really go after it and use it like mad. However, subsequent work that Desmond Cook did, I guess, wasn't very successful. But he said we worry so much about it and think it's omnipresent, so maybe it's something better than most of the treatments we can think up, and why don't we really go after it and use it like crazy...

**Hagin:** Can you recall, John, Desmond Cook's findings? I heard his description of the initial design, but I've lost track of it.

**Feldhusen:** It's so far back now...I simply recall several conversations which suggest that the results were that, in whatever areas he was investigating, there wasn't much variance attributable to the Hawthorne Effect after all.

**Millman:** That particular study of the Hawthorne Effect is flawed and should not be taken seriously. But, you're right, it's conclusion, which I don't feel is justified, is that there wasn't much of a Hawthorne for that particular situation. The thing to remember about a Hawthorne Effect is that it is an ambiguous term and it's been given many meanings. One student wrote a paper on that and identified four different ways in which it's been used. In general, it's attributing the superiority of one treatment to something other than its nature because of something this is considered extraneous to the treatment, such as special attention, or awareness that you are in an experiment. The thing to remember, as John pointed out, is that there was a superiority and if, indeed, being aware that you're in an experiment, or whatever caused it, then why don't you just incorporate that into the treatment? Another way to put it, I think, is that a Hawthorne Effect is still an effect, regardless of whether or not you consider it contamination of a sort. The superiority of the one group is still there and it is a real superiority. If novelty or some of these other ingredients were really responsible, then maybe we should just try to incorporate in the treatment that feeling of special attention—that feeling of being aware of being in something rather special and new.

**Proger:** Do you have any further ideas on the Hawthorne Effect, or it's control?

**Millman:** Well, yes. You like to be able to pin down the reasons for something being a success. Clearly, if you have confounding, in that the one group not only is getting their special training or whatever, but is also getting a feeling of being an experimental group and getting special attention and so on, it's nice to be able to say which of these things was responsible. If the two are confounded and mixed up, you can't necessarily separate them out. If it's important for you to separate them out, then you need a control group that gets special attention and is made aware that they're something kind of special and given something kind of new. In other words, to have everything but the treatment.

Feldhusen: I have a feeling, like Glenn, that a treatment should undergo a programmatic round of evaluations. Now, it does seem to me, that with a treatment: in the first round one might have a rather complex, undifferentiated package of some kind or other, or set of instructions that constitutes his treatment, and that package might include Hawthorne Effect, too. It is not unreasonable to go through the first round being delighted if you simply get a significant effect, regardless of reasons. The, in subsequent stages, we become concerned not only about sorting out this so-called Hawthorne Effect—if it's a part of your package—but the treatment effect, too. The developer, indeed, may want to know what are the things within his package. Art Lumsdaine first discussed this in his chapter in the Handbook of Research on Teaching. It has since intrigued me many times. How could I design my research now, knowing that there's the possibility that my total package works. I still strongly suspect that parts of my package are maybe just dead weight that goes along without doing anything for me. So, I would like to design a study, or a series of studies, which begin to break the package down, to determine what it is that is really helping to produce this gross effect that I thought I was getting at the beginning, here. So, at that stage, one might become most concerned about sorting out the Hawthorne Effect component. But, for the first round, let us accept the Hawthorne Effect and any other components, and simply see: Can we get an effect with the total package? If we can't, with the total package, then I'm quite negative about going on with any further research.

French: Let me interject that I think this concept is something is something that we need to get back to when we talk about diagnostic-prescriptive teaching, and saml1 modules, and four year programs, and so on. The things John was referring to seem applicable.

Hagin: Can we turn to the topic of retrospective control groups? For example, children of the previous grade, in a traditional program, were used as a control group for an experimental intervention with the current year's group of children. I had been criticized for this in a paper I wrote.

Feldhusen: From what Rosa described, it seemed that a group of critics, hearing her paper and having found a flaw, then exhibited a behavior that is pretty characteristic, that is: Your study is out, bad, totally; we found a flaw that is on the list for fourteen threats to internal or external validities. Then that led me to ask Jay, since I know how much he is concerned about the evaluation of research, "Has anyone thought seriously about devising ways to assess the amount of loss in inferential values in each one of these criticisms that the critic finds?" The journal editor goes down the list, and there is this threat, and finally he reaches some threshold where he rejects the manuscript. But we agreed that this was a very subjective process. How much nicer if someone were, somehow or other, beginning to try to quantify the degree of loss for each of the criticisms.

Millman: My answer is: you arrange for other control groups to yield additional information, so that you can tease out what seems to be the contribution due to certain things. I think, that in a paper in AERJ four years ago, Gene Glass actually gave a paradigm for doing this in one case, where he showed how by getting additional control groups that have this and not that, etc., you could tease out certain things.

Stanley: A very nice paper. I just recalled that years ago I was consulting with some team-teaching people at Racine, Wisconsin, and one of the schools took three teachers: history teacher, English teacher, and math teacher. They wanted to teack the 90/honors people who could schedule history, English, and math honors courses. They wanted to teach them as a team; they wanted to evaluate this effort. This was not a social studies team, but was a horizontal team for a particular grade. We speculated that we could have various types of controls; we could have three different ones, and see if we could get different results. No one, of the three, would be very good. One would be to compare the students that could schedule it with those who couldn't, using some kind of covariance adjustment, perhaps. Another would be to look at last year's group, where there was no such program going. Third, go across town and get another high school that didn't have this program and take some baseline there, too. And then, the idea was that if these three all agreed, since they had different probable sources of bias, then you could probably say that you have a pretty tight control, even though you don't have random sampling. Unfortunately, since I simply talked about this with people, it never did get implemented. It is always a question about the various sources of biases that could occur with last year's classes versus this year's—the change in people year-to-year and, of course, the changeover time. I guess there is never any definitive answer.

Hagin: We just used the scores.

Stanley: The direct score.

Hagin: At the completion of second grade.

Stanley: You didn't have...then how do you know—this is the question—How do you know that the children started at the same point? Did they have different teachers or something?

Hagin: All of them came from the same school. They had somewhat similar teachers.

Stanley: Two successive years? How large were the groups? The whole school population?

Hagin: Between 80 and 100 in an entire grade...

Stanley: Everybody in the same grade each year. How were their neighborhood changes in the neighborhood structure from year to year? You see, there are all kinds of things you can think of, such as—no one of them seems, in itself, important. Were there migrations of a particular minority group with the class regulars?—something like that. From one year to the next, there might be a gradual attrition, that is, moving to the suburbs, and that could affect the level from one year to the next appreciably.

Hagin: No, I was not displeased with this, but there was criticism on other grounds. There were no marked changes in the neighborhood—no urban reason for this.

Stanley: One tactic to use with your critic is to say, "What do you think the sources of perturbation were?" Then, take a while for them to toss in that, let's say, the Lebanese did move in, and so forth. You go along, if you are doing something ala B. F. Skinner, where a huge difference is going to be manifest. Then, probably, the differences from year to year aren't important. You know, Skinner has never had to have control groups to use statistics because he has refused to work with variables that don't make big differences. When you change the rat's ration, you expect every rat to do this, and if every rat doesn't do that well, Skinner doesn't play that kind of game. He gets himself another kind of variable that is more important. But, in the situation where there is a moderate signal, and a lot of noise, whereas not a very important difference not likely to be elicited, and if you went all out with everything you had all year long, then maybe you could ignore the difference, because you could expect a huge gain—a Big difference—and much of that would be real, even it there had been some change in the group from one year to the last. So, I would think how hard you worked to get gains, and how confident you were that big gains would be yielded.

Mann: Ken?

Hopkins: Glenn and I employed two of those three types of control groups. We were doing some work at Colorado Springs and a neighboring junior high school that seemed to be similar in socio-economic status. Last year's results of students in the same school, same grade level, were from some aptitude scores obtained a few years earlier.

Stanley: Did you get consistent results—both ways—when you got the same results? I think, that would add strength to your feeling, that it was correct if you got one set in one direction and one set in the other direction, you would really be in an interpretive dilemma. You all know by now, thanks to Fred Lloyd particularly, that the massive covariance doesn't do magic things for us, and yet, we are probably better off to use it in some of these situations than not, particularly when we made every effort to get a comparable school across town and comparable measures...probably help some to equate the groups. The big question is how big the difference is. It is a question indirectly of power. It is not quite power, but it is related to power, because you are always concerned with how big the difference elicited will be, and, therefore, how hard it is to detect, or how easy it is to detect, and how big an experiment. In this situation, you are concerned with how big the difference is, relative to the disturbing effects, themselves. One such effect is the sampling variation, of course, but you also want to know what disturbing effects introduce some kind of systematic change. Now, what about that—where you have the neighborhoods of whites that are fleeing to the suburbs? Say, fast...That from one year to the next there could be a marked change. You can predict which way it can go. If you had initial measures on both years, you would know the size of the treatment effect. I think, that the gain measure procedure—pre-test/post-test—is almost essential, if one uses another year for comparison. You need to know the groups were, to begin with—because you really have an interaction question—if the gains one year are greater than another year, you really have differences—a reason—than if the in point each year is the question of where the beginning point was left undefined. You really have a question, don't you, Les, on what you are going to use those gains for. If you are just comparing average gain of one group with average gain of another, that's quite a different situation than, say, trying to get gains for correlation purposes, or make comparisons of two individuals in gain score, or where you have a selected group when you are trying to—when you have the tail of the group—when you have the regression problem. The simplest problem would seem to be when you are trying to compare the mean gain with one group with the mean gain of another—an entire group—whole groups. That doesn't seem to be such a big problem compared with this correlation of the individual assessment—individual differences problem. Had you gone beyond the problem? Come back in phobia type.

## REFERENCES

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.) Handbook of research on teaching. Chicago: Rand McNally, 1963, pp. 171-246.

Clasen, R. E. Book reviews: An introduction to educational research, third edition, by Robert M. W. Travers. Journal of Educational Research, 1972, 66, 56 + 60.

Gage, N. L. Paradigms for research on teaching. In N. L. Gage (Ed.) Handbook of research on teaching. Chicago: Rand McNally, 1963, pp. 94-141.

Gottman, J. M. N-of-one and N-of-two research in psychotherapy. Pscyhological Bulletin, 1973, 80, 93-105.

Poulton, E. C. Unwanted range effects from using within-subject experimental designs. Psychological Bulletin, 1973, 80, 113-121.

## CHAPTER 17

# RELIABILITY IN SPECIAL EDUCATION MEASURES, ERROR OF MEASUREMENT, REPRESENTATIVE SCORES, AND CONFIDENCE INTERVALS

Overview: A case is made for the more frequent occurrence of unreliable tests and measures in special education than in regular education. While one is certainly concerned about reliability of his tests for individual decision making, unreliability is not a major problem when comparing groups of children; thus, while one always strives to get the most reliable measure for a particular task at hand, he need not worry unduly about the effect of mediocre, or poor, reliability on results of experimental research. However, there is one notable exception in experimental or quasi-experimental research, where unreliability indeed hinders sound analysis: analysis of covariance, where initial differences are due to lack of randomization and miscellaneous biases. In effect, because of the well-known equation, i.e., the variance of observed scores on a test is equal to the variance of the true scores plus the variance of the error scores, unreliability increases the variability of observed scores (which are used in analyses) and thus, results in a less efficient analysis of the estimated true relationships among variables or factors. But the problem is generally not severe. On the other hand, when one turns from experimental research to status studies (such as are embodied in multiple regression), unreliability of the criterion variable can lead to real difficulties. Miscellaneous points are also raised in this chapter about covariance, size of sample, test selection, and norming efforts.

Cross: Reflecting back on some of the things that were said this afternoon, in terms of experimental designs and what not, it just occurs to me that the questions of whether or not an N small is troublesome, or whether you believe or don't believe in randomization, taking out covariates, etc., all of this is secondary to another problem that is, the standard error of measurement in the instrument you are using. The power of your statistical tests, etc., only considers sampling variance from the population, and the actual measurements you are entering with are infallible. It seems that, particularly in special education, you are dealing with a lot of instruments which are not reliable, which have horrendous standard errors of measurement. For instance, any one of the subtests on the WISC have reliabilities which are rather limited. The Bender and a lot of other psychological tests, as opposed to achievement tests, are rather limited in their standard errors of measurement. These studies in the Journal of Special Education, for example, use these tests and then do an experimental design with a test for a statistical inference. They find it to be significant, but I'm worried about the fallibility of the data they entered into it. Is this a greater problem for special education that deals with limited numbers and with fallible data, than is the case with regular education?

Hopkins: Well, I don't think there is any assumption that says analysis of variance demands the criterion variable be measured without error. It is true that error measurement is going to increase the error variance, or the error term, and make your test less powerful than it would be if the criterion were more reliable. I think the real issue comes into the picture in analysis of covariance, when you are in a quasi-experimental situation. That is, you have not been able to randomly assign subjects to the treatment groups and initial differences are no doubt present between groups. Your statistical compensation is not going to be perfect. Besides, covariant is pretty unreliable, to the extent...if you really haven't taken out the bias. On the other hand, if it is a true experimental situation with randomization present, then I don't feel it is a serious problem, even if you covary an unreliable measure, because you are not going to be adjusting the means that much anyway. So, in other words, I don't think it is anything like a violation of a statistical assumption in analysis of variance. It is a matter of efficiency, and we get little more powerful analysis if we have a more reliable dependent variable. It becomes critical when we have large initial differences that we are trying to statistically compensate by analysis of covariance.

Millman: Well, that is pretty much what I would say. As I read it, a few additional points might be made. First, the problem of using unreliable measuring instruments is no problem, if you are talking in terms of scores in the test. In other words, you want to make statements about the superior performance of group A as with group B on the WISC subtest 3. If you want to make inferences about the true scores, so to speak, and where they're real scores—which I think we most often do—then there is no bias by using unreliable tests. The problem comes in your being more likely to make a type 1 error, mainly; your type 2 error, you are less likely to make. That is, you are not as likely to claim statistical significance—but, perhaps, you should—if you use an unreliable test. The problem is more acute when you have samll N's, than when you have larger N's, as I recall. The final point that might be made is that if the test is not valid, as well as not reliable, then there is no way out.

Cross: Do you find the best estimate of group's true score would be the observed group mean? And are you saying that, because of analysis variance and factoral designs and that sort, you are using the group mean; you are saying that it is less affected by unreliability and therefore, it would not cause any great problems?

Millman: The best estimate of the true score is not the observed score.

Cross: The mean group score. I am not talking about an individual. I am talking about a group.

Millman: Well, as Ken said, the problem is that it decreases the precision of the test. It just makes the error value bigger than it really should be. But, I think, as I recall some papers I read on this, that the problem is less severe as sample size goes up and the reliability obviously goes up, too. But, if you have unreliable measures used with samll r studies, there is a problem, but the problem is not one of bias.

Cross: It is one of efficiency.

Millman: Well, without the covariance, it is probably a lack of precision. But people use analysis of covariance to try to hold constant or correct for initial group differences, and you are undercorrected when you use fallible covariants—just as you are undercorrected when you don't use a covariance.

Bracht: Another way to express this issue is: the variance of observed scores is equal to the variance of the true scores plus the variance of error scores. The less reliable the measure, the greater the variance of the error scores, and that increases the variance of observed scores which is what we use in the analysis of variance. And so, to the extent that we have less reliability, we decrease the efficiency of the analysis.

Feldhusen: Let us simply remind ourselves that in multiple regression analysis, we would not expect greater predictive efficiency than is inherent in the reliability of the criterion. It seems, sometimes, people need to be reminded of that. One cannot expect a greater predictive efficiency, multiple RR squares, than the reliability of the criteria that he is attempting to predict. In Lavin's review of prediction research, he complains that we have gotten characteristically up to multiple R's of .60, .65, etc., and he suggests that there must be great room for improvement if we will look to personality assessments, etc. He neglects to make clear that the upper limit of predictive efficiency might well be chiefly present in the kind of academic criteria that is being predicted and that, indeed, those multiple R's at .65 or whatever it is, may, indeed, be the upper limit for the present, in that no amount of addition of personality, or attitude, or other variables, will increase that predictive efficiency until there is an increase in the reliability of the criterion.

Hopkins: I think it is a very good point, and I think it is also the chief reason that the efforts to predict teaching success has been so poor. It may not at all be that the predictors are not relevant. It may be just that the criteria that have been used have been so unreliable that nothing could predict them, regardless of how valid predictions were in the labs.

Millman: Then, one could look at the other end at the predictors, and if you've got error, you know, from unreliable predictors, then they won't be as successful in predicting the criterion. That is, if you were to say, use longer tests and more reliable predictors, that should increase the predictive validity of a particular subject. Not too much, however, perhaps because it isn't a linear thing. If you increase the reliability from, you know—if you double the reliability, you are not going to double the predictive validity.

Bracht: Regarding the point John made, if we obtain a high score which is greater than the reliability of the criterion, is that due to...well, is that something that couldn't be cross validated then, or could that occur, you know, within a sample of data that we might obtain an R square radial reliability, and that would just be something that could not be cross validated. It would be an artifact of that particular sample of data.

Feldhusen: Mr. Ishmael at Purdue had some good examples. The first time I ever saw multiple R's of .95—I believe for some achievement criteria for which we ascertained that the reliability of the test didn't affect it—I don't think he was fallible in sample size either. I guess he just happened, by chance, to pick it up.

Bracht: He had so many predictors that he capitalized on some, correlated with the error scores.

Hopkins: I think there are two factors involved here. One is that, of course, multiple R is always an inflated estimate, and it is inflated in direct proportion to the number of predictors over the number of subjects involved, and that degree of inflation becomes sizeable. I remember, in some of my work, a multiple R of .6 with about 20 predictors on 60 people strength; that is how much inflation comes into the system because of capitalizing on chance error. The other thing is, of course, it is technically possible to predict a criterion, although not the upper limits. This deals with the squaring rules of reliability for the criterion which is greater than the reliability itself. It is very unlikely, but theoretically possible, to have more validity than the criterion limits reliability, but it is just not going to happen.

Bracht: It is assuming that the predictors are perfectly reliable.

Hopkins: That's right. I get a parallel test, twice as long, as a predictor. It will predict the criterion provided the reliability of that predictor is perfect, but that is a very artificial situation.

Cross: Well, I just wondered if that same thing that you mentioned, where your predictors are large and your N is relatively limited, and then in cross validation, your multiple R shrinks back to 0, which is the same thing, I guess. Cureton explained this very well, in his article. Couldn't it be the case, that we are capitalizing on the errors of measurement in predicting, and also, where the scores that we plug into an experimental design have experimental measurement error in them? In these two cases, we are capitalizing, perhaps, in a different way—it is not the same way, obviously. But can it be that, in some instances, like: they did a study once and it is not significant; they do it again and it is a significant one, so they send in the significant one for publication, and it may be that that significant one is attributable to the errors of measurement, or whatever way you care to look at it. You can play the odds that way.

Mann: Well, if you repeat your study enough times you are going to get...

Cross: But, by and large, research is usually a one-shot story, and I am saying, he is likely, as not, going to get significant results, as not significant results, partially attributable to the errors of measurement, whether positive, or negative, and so one. I guess what we are saying is, over the whole, positive errors of measurements are going to cancel out the negative errors of measurements, and some spurious significance is the consequence.

Hopkins: I think the multiple R's are inflative, even if your predictors have no measurement error, now. And the example I gave was not an empirical cross validation. It was simple a theoretical statistical crutch, a procedure for estimating what the unbiased multiple R is when you feed in the parameters of number of predictors versus the number of subjects. In other words, it is a prediction of what would happen if I had cross validated. I didn't cross validate to get the multiple R.

Proger: How about the issue Larny raised about the positive results due to the error of measurement?

Millman: Well, on the whole, it probably doesn't work that way...

Hopkins: I'll speak on the opposite side of the fence. I think that, often, there is more to be gained by, let's say, cutting a criterion in half and paying the price, in terms of power, by being able to get a measure of another important outcome variable, in the line of what Glenn was suggesting earlier. I would strongly prefer tests with moderately reliable criteria, but several of them, versus that single, two-hour measure of reading vocabulary with a .98.

Millman: When you have a prediction system, I think it behooves you to cross validate, or otherwise try to re-validate your work. Editors should insist upon cross validation, also, while there are no hard, fast rules, one rule of thumb that makes some sense is that if you have, say more than 3 predictors—3 or more predictors—then your system probably ought to have something in the order of roughly 20 people per predictor. You know, if you are studying 5 predictors; probably shouldn't even try it unless you have 20. The two studies that I know of, seem to come out with something like roughly 20 cases per variable. This minimum makes it kind of tough, when you are dealing with 10 subjects on a research. The trouble is, if you don't observe the minimum, chances are you are going to pick up something that will wash out when you try prediction at a later time.

Feldhusen: I simply want to echo what Jay has just said, in relation to the concern about small _N_'s, that it is, indeed, in this domain of regression analysis, that there simply is no way without quite substantial _N_'s...practically no talking at all about studies involving 10 subjects, or 7, or what have you. Although his criterion of 20 to 1 is a bit more severe than I've heard as minimums, at least I have never heard anyone suggest a criterion of less than 5 to 1, and...

Proger: I have heard of 30 to 1.

Feldhusen: But, you see, even 5 to 1, it is a rare multiple regression study that doesn't have more than 3 variables involved, too, and multiplying even starting at the level of 5 would quickly add up to quite substantial numbers.

French: I'd like to go back to the previous issue for an interlude. One was just dealt with—the importance of determing reliability on the instruments that are used in the study at the time of their use, and how important the group feels that might be. Let's take a test that has reliability data for subtest reporting on standardization sample. Yet, in many instances, the kids that are being tested in special education were not really represented in the standardization of the test. How important is it to determine reliability data when the group you are working with wasn't represented in the sample?

Mann: Ken?

Hopkins: Well, I personally feel that, in most instances, reliability is only a secondary interest in an experimental design contest, as long as the information is not used as a basis for individual decision making. I think, it is always good and nice, to have all groups represented in the normal group, but of course, what you work with never will be the same as it would be in the normative sample, even if you are working with normals, because anytime you change the variability of the group you have, you are going to change the reliability. If you have a more homogeneous group, you will have a less reliable scale, whether MR's are represented or not. I think I would say, desirable but not essential, as concerns your point, Joe, in terms of use for individual assessment—reliability, of course, has to be very high.

Mann: Jay?

Millman: I see you as having 2 questions, Joe. One, how important is it to look at the reliability separately for special subgroups, and is it important, anyway, in research! First question: I would say it is important to re-compute it, if your sample will not provide the same range of performance as the other group. If your sample is going to be at the low end of the scale, or something, then I think it is important to re-compute it for the reason that Ken gave. I think what Ken has in mind, regarding the importance of the reliability data in research, is that the: what you are concerned with, is not so much the reliability of individual scores, as the reliability of the group performance and the reliability would be much more stable or much higher than the reliability of the individual scores, so that the pluses and minuses sort of balance out. However, where reliability gets important is to help you interpret your results. If you don't get your treatment working out, then your next question is: Why isn't it working? There are several reasons: one, it is too small a sample or something; another might be that the treatment really wasn't very effective and it just isn't a good treatment; but, there is another reason, and that is, that the instrument that you are using to measure treatment difference might be unreliable. If you are dealing with an unreliable instrument, then no matter how effective the treatment is, if all your measuring with your dependent variable is error, then you are going to get null results...you are going to get an expected difference of zero, and if your instrument is very low on liability, and your sample is somewhat small, then you run the risk of failing to find differences, when you could have had a better instrument. I'd say, the importance of reliability is increased when you fail to find results. If you find results, in spite of the possibility of having an unreliable instrument, then you don't worry by not having that reliability data. But, if you don't get positive results, then you look around for the reasons, and one reason may be unreliability and that is why I think, in that situation, it is important.

Bracht: This is not directly related to this, but I wish researchers would be as diligent in studying the validity of the instruments, as they seem to be in studying the reliability. It is so much easier to study the reliability and, again, I think...you know, that is secondary to validity.

Mann: I think that the untrained test user confuses the two, rather, sort of a cognitive blue. He just reads by and sees the figure—there is the .85—and off he goes. I think, typically, the teacher does this and doesn't look for the validity coefficients which usually aren't there, anyway, in most of the new tests that are being published in special education. Right, Bart?

Millman: What, other than achievement kinds of tests, what kind of measures are special educators interested in, as criteria? What are they interested in, besides the achievement, tying shoelaces?

Hagin: Well, here is one to think about that really gives me nightmares: defining social maturity which has been around for so many years, and is frequently used as an ego factor, and other personality variables. And yet, what else? Measures of language—hard measures of language—we've been playing around with these for so long...pre-school achievement. Most of such scores now offered are mishmash affairs composed of borrowed items. Poor Binet!

Bracht: To what extent, in special education, do we like the criterion variables to be the performance measures rather than paper and pencil measures?

Mann: Well for one, in vocational. But the success is limited.

Feldhusen: The Belmont project office of education certainly has made one massive effort to get at criteria and measurement problems, particularly in relation to assessment of disadvantaged youngsters, if I remember correctly. Belmont was specifically concerned with developing, or modifying, or renewing, and so forth, of tests to make them more "useful," meaning, hopefully, more reliable and valid. That was a large cooperative venture—Project Follow Through—and, also, has attempted to deal specifically, again, with the problem of criterion measures.

Mann: Are these performance measures, John?

Feldhusen: I think the Project Follow Through—I went back to ask what are the kinds of goals, or ends, that characterize the various programs and then what are existing instruments, what are instruments that need to be developed, to do a more adequate job of assessment? In general, what I am saying, here, is that at least in two instances, massive efforts have been made to try to attack a problem systematically and on a large scale. I had a part in the Project Follow Through, and was much impressed with their effort to try to survey existing instruments that they might be neglecting, or that they might be drawing in for use; mostly paper and pencils, to be sure, and then to go on...to identify what are the critical areas in which we need to see if we can stimulate test development improvement, or the development of new measurement procedures. A worthwhile undertaking.

Mann: I just jotted down various topics, after speaking to various people, and would like to pass around the table, about the specific to special education. Now, a lot of the issues here are not certainly limited to special education, but they are important for special educators to hear and to discuss, because they rarely ever do. Larny, for one, felt that as far as he was concerned as a research intern, that the issue of a reliability had not fully been cleared up and that was particularly important to him, in terms of what was the issue we discussed yesterday...measurement of change. Would you want to bring up your questions about reliability one more time?

Cross: One of the points that I didn't emphasize, I guess, and that was, that with special education we tend to, oftentimes with instruments, which are projective or which are not highly reliable and where the criteria is not unitary, it's a construct you are measuring as opposed to specific criteria and then you look at the projective test, whether it be the Rorschach or TAT, or if you will, the part scores on a WISC, or other measure of intelligence. Those are based on norms of a normal population. We not only have an attenuation in the variability of our sub-sample, but they also happen to be at the very lower end of the continuum, where, it seems to me, the error of measurement is even greater. And to block on a variable like that, or to use it as a covariate and set up your experimental designs where your reliability goes down almost to nothing effectively, you don't know unless you investigate it on small N. It seems to me, it is like garbage in/garbage out, in a lot of these things, and it is of particular relevance when you are dealing with 10 subject, or so, in special education, when you block on a very unreliable instrument or covariant. Soooo...can we dismiss the reliability again as we talked about, particularly the types of tests used in special education and for the sample sizes we deal with?

Mann: Everyone has been rather casual about the reliability—about it being critical—you need a more powerful test, perhaps. You are going to lose a little power, but, by and large, it has no concern with a regular research design; that's important for individual cases.

Hopkins: Well, we did say, I think, it was critical where you are having your lion and mouse covariance, to remove bias.

Mann: Yes.

Hopkins: On the other hand, let's take the degenerate case. Let's suppose you assign random numbers, as a covariate, to see what happens to the general reliability. You are really going to suffer very little, in terms of a true experiment where the covariate has zero reliability—random numbers. You run the analysis covariate. What it simply means is, we haven't reduced error variance at all. Our adjusted means will equal our unadjusted means. We have lost a degree of freedom, but really, we haven't noticeable suffered anything by it.

Feldhusen: Again, quite a different problem was put to me in one of our discussions. First of all, since we are talking about a covariance situation, let us think of the correlation of the covariate with the criterion, and consider the possibility that in a very small sample that correlation may run spurously high. Now, what we have more commonly worried about, in a covariance analysis and, partly, what your answer relates to is correction because of lack of reliability in our covariate, we do not correct as much as we ought to for our potential bias among sub-groups. Now, I think it is rare, but yet, I think it is possible, that because a very small sample which breeds a spurously high relationship between the covariate and the criterion that your would instead get a, let's say, a spurously high correction occurring in a covariate situation. What is your reaction to that possibility?

Hopkins: Well, I recall, that it would have to be the same kind of spurious correlation within each group. That is, it isn't just the overall correlation; that is immaterial. It is kind of the average correlation within each group. Yes, it is theoretically possible, but let's say, in five experimental subjects, you have a very high positive correlation, and in your five control subjects also, but it is most unlikely that you are going to get the same degree of extreme error in both independent groups. It is kind of like the average correlation within the smallest subdivision determines, you know, the average slope which has been too much to adjust, both the means, and to get your air trip. So, I guess, it is theoretically possible, but...

French: Will this position be modified somewhat by the characteristics of a particular sample, and I guess what I mean by that is, that many times your reliability coefficients are determined for one age level, or one age grouping—for 6 year old children, or for 7 year old children. They have two reliability coefficients for 6 year olds and another for 7 year olds, but in many special education classes, there is a range of ability there, back to the small group situation where you have 10 kids, maybe, spread out over 5 ability levels, and so when used with those particular kids, the likelihood of ranking them in a different order on a second administration may be unlikely, even though the reliability for one age level might suggest that that would be the case.

Hopkins: Well, I think, the notion of correlation, in terms of dealing with the relationship between covariant and dependent variable, is strictly a heuristic mechanism. As a matter of fact, it isn't a correlation at all...it is the slope and the variability; if we had gotten any averages is immaterial, in terms of changing a slope; it will change the correlation but it won't necessarily change the slope. Although, I think, very good point can be made about how little we know about what happens to that regression line way out in those trails. I think, in certain instances, I tend to disagree with the point that as you get more deviant from the mean, your error of measurement builds up. We know that on the Binet, that certainly, the reverse case occurs with big heartaches. The error of measurement is less for 70 IQ than for 100. That is, that inverted pair phenomenon occurs, but I think in most pairs...I don't know the literature, but I would suspect we know very little about what the aggression means level off and so any projection or extrapolations, I would think, are very suspect.

Millman: Well, Lester, I don't think we really dismissed reliability. I'd say that we recognize that, with deviant populations, instruments may be, indeed, very unreliable. I think we also recognize that this was a concern, especially when looking at individuals for purposes of making individual assessments. What we did say was that in the case of research purposes, in looking at groups, that the unreliability should not introduce bias into the comparisons we are studying. It may produce an undercorrection. It will interfere with the power of your significance test, but you are not introducing bias in an actual experiment. The real concern, or what I call the one concern, with reliable instruments for research purposes, is this lack of reduction in power which prevents you, frequently, from recognizing potent treatments or real treatment differences, when they are there, because your instrument is not good enough to pick it up.

Mann: Rosa? You are out there doing a lot of this research with these deviant youngsters. What are some of your particular problems?

Hagin: Well, I am concerned with achievement batteries as they are administrated. You have no idea why the child is responding in the fashion that he does. Now, we do have information from the technical manuals that inner city children of the large cities have never been utilized in such standardizations, and furthermore, if you look at one child and the test, you will find that they respond in ways that are really not respected by the people who made the test. It is a very serious question, because we are not at all convinced they offer a valid measure. Now, there are other problems; i.e., overage and underage children are omitted from this standardization of the Metropolitan, for example, and there are real questions of whether the contents of the Metropolitan is appropriate with our sensitive children, and we have done item analysis and found very peculiar results with it. Other people have similarly found it for the reading tests.

Proger: We have been concerned, in our recent research, with the unreliability of physical performance increases. If you have a specific measure of knee squat or, for example, the kid bends down wrongly, there he goes, the score is shot as far as being representative of him! Handicapped youngsters, there is just no doubt, are extremely more variable from trial-to-trial. P.E. people have handled it in a number of ways. One, is to average the number of trials, or you count the best of the three as being representative of performance.

Hopkins: Well, of course, one way of improving reliability is to increase the number of trials and average those, and so, I think that seems to me, a very reasonable strategy and if it isn't reliable enough, well then 3 trials and then 6 and so forth. I would think, from what I know about it, that that would be preferable to taking the best, which would emphasize extreme unrepresentative measurements.

Bracht: I don't know how reliable this procedure is, but in gymnastics meets there are judges. The judges who give the highest and lowest ratings are not counted in giving scores. It is the two in the middle who are averaged, assuming that, you know, you may have one who is deviant in each direction, and the two in the center are average, but I don't know if that would provide any more valid or reliable scoring procedure.

French: I think it also depends on what you are looking for, in the way of a test score. Are you looking for an indication of one's maximum ability to perform, or his typical ability to perform.

Mann: For estimates of group, rather than individual, averages, would the group be a more effective way of increasing power here, or would you still recommend the averaging out in physical performance situations?

Hopkins: Of course, increasing the N will increase the reliability of the mean. It will not increase the reliability of the measure. It will give you a more precise estimate of what the reliability is...a small confidence internal, but...it will not increase the reliability; just give you a better estimate of it.

Feldhusen: I would like to go back to your original assertion here and raise questions about why this area of measurement is characterized by a dramatic lack of reliability. I wonder, is it all a product of poorly defined measurement procedure because if it does, I am going to pick the one example I know well, namely, Mr. Ishmael's work with retarded children. I recall they report quite good—a high reliability for those that have a large battery of physical assessments, many of which are very complicated, you know, from walking the beam to all kinds of jumping, hopping, and various other things.

Mann: There is a reason for that, John. They work theirs with EMR youngsters, first of all, and I think that EMR youngsters, in many instances, especially inner city youngsters who may be quasi only...these youngsters do quite well in motor skills and, in fact, better than "normal" youngsters, in a number of instances. But, if you work with trainables, severely brain injured, or emotionally disturbed, there is erratic functioning from trail-to-trial if you often don't know what the child is responding to, when you give it directions even of a very simple sort.

Feldhusen: This is accounted for in the Ishmael measures. In fact, that is one of the criticisms of his work. He gets started in high correlations between his physical measures, and achievement, and IQ assessments and the standard criticism has been that there has been so much mental activity in trying to understand the directions of his task, that it is inevitable that they would correlate with IQ. Nevertheless, he is able to derive a measurement—a score out of things—through his observation that it is accurate and reliable and the fact that you can't...Now again, I can't say that he has worked with trainables. I don't believe they have, but if the...

potentially, if their observation procedures are very well developed and specified, then the poorer performance of the trainables should not, in itself, make his observation procedure less reliable.

Proger: I do think the evidence is that they are a much more unstable population, as far as physical performance increases, are concerned. It is a matter of performance, perhaps, rather than ability, which fluctuates.

Hagin: Are we saying then, that increasing the number of judgments is better than increasing the size of the population studied on such measures? It has implications, in terms of neurological...

Millman: Well, I answer the way Ken did. If you are interested in making judgments about people—individuals—for their own future diagnosis, then, obviously, you want to increase the number of trials for them to get at, as reliable measurements of a particular skill.

Feldhusen: I think you are suggesting the poorest solution here. The best solution is to improve your methods of measurement. If that is not possible, then I'll agree with you, but still my assertion is that first, the critical, most important issue is to look to your measurement procedures, to see if they cannot be improved, and your solutions I relegate to the last resort domain, if you wish.

Millman: As I see it, if you are interested in a person's ability to walk a bar, or whatever, the question is—if you are interested in that—are you better off to use more people with fewer trials, or fewer people with more trials? In case you want to make statements about individuals diagnostically for people you, obviously, can have as many trials as seems to be needed, depending on how much variability there is. In the research context, I think it generally turns out that you are better off having one trial. I think you gain in introducing sampling variability, generally.

French: Particularly when looking at the emotionally disturbed and neurotically impaired and when trying to determine the maximum ability that the individual has, I would think it would be desirable to have more trials on different days, in an attempt to capture the maximum performance that individual is capable of, as opposed to just sampling him one. I guess, if we had a lot of cases, then the average would work out but I think with people who have a motivational problem, I would feel more confident if we sampled them on different days under somewhat different conditions.

Millman: Well, it depends a little bit on the question you are asking. You are asking: Can this group of people, called the experimental group, walk the bar better than the groups that we call the control group? Then, I would think I disagree with the saying: that you just increase the size of your samples or re-plot trials.

French: Sometimes, there are academic predictions studies, based on physiological performances, and then it is an attempt to get a capacity measure, rather than to select whatever the kid was able to do on a particular day.

Millman: Oh well, in prediction studies, that may be different. Let me think about that. That may be a different...if you are trying to estimate, not whether this group can walk the bar better than that group, but what is the real relationship between walking the bar and being able to do well on some cognitive task. Well then, I think what you want, obviously, is you have to have certain degrees of reliability in your measurement. If you don't, all the cases in the world aren't going to help you. That's a different story.

Mann: Glenn?

Bracht: If the thing that we are measuring is highly variable, within the individual, I think that is important information that we do not want to mask by just giving a large number of trials and taking the average, because just taking the average is concerning a lot of important information and we somehow have to be aware of that, and treat that into our analysis. I think if we have five trials for a person and all of those five observations are very close together, that has much different implications than if we have five trials where the score is spread out to a large degree.

Mann: How would you approach it?

Bracht: I am not sure except, I think, in the average, it may be a misleading score. If the scores are close together, the average is representative of that person, but if they are highly variable, the average is not a very revealing score or a representative score for that person.

Feldhusen: I stand by what I said before.

Hopkins: Again, just to go over a point that I think Jay and I have been talking about, I think it all depends on the purpose of the information. I think, in a multiple regression, or prediction context, reliability takes on a completely different complexion than it does in an experimental context. If you have a reliability of 1.Q and a small *N* you still have a very unreliable estimate of the population norm. If you have to choose the lesser between two evils, have a less reliable measure and a larger *N*. You'll get a better estimate of the parameter and a very small number of the 1.Q. reliable. I think it all depends upon the purpose. For individual decision making, of course, there is just no question you need all the reliability that you can get your hands on and that is what you are doing in a prediction context.

Millman: Lurking behind all this, is the whole question of validity. For questions like: Should standardized tests normed with "normal populations" be used in studies on special education?

Mann: May we discuss the issue of confidence intervals in a very simplistic way, since there is no doubt that it is a constantly confused and misinterpreted issue.

Hopkins: Well, I don't think it is important, really, that it is misinterpreted. I think the spirit of interpretation is right, even if it isn't interpreted correctly—to meet the letter of the law, so to speak. I think the contribution it has to research is related to the problems we were discussing so much yesterday, about small N's. I think, if researchers reported confidence intervals, rather than just acting on the null, we would be far better off because people would be much less inclined to believe in Ho when it has not been rejected, if they saw that still the correlation might be as much as .6, or whatever, even though it wasn't large enough to reject the null. I think confidence intervals become especially important in small N studies, because they force the reader to think in terms of the degree of error in the estimates that are arrived at. In other words, treatment A is much better than treatment B, even though the difference wasn't large enough, in this study, to reject the Ho's. So, I think confidence intervals have a special importance, goal, in small N study.

Bracht: I think confidence intervals also facilitate comparing the results from one study to another. I was assisting a student last year in the counseling area, but he was administering a treatment for which, in a number of studies, the same dependent variable had been used. And in talking about these previous studies, he always had reported if the hypotheses were accepted or rejected and this really didn't provide very useful information, because although the experimental groups tend to have some commonality across studies, the control groups tended to have different characteristics which could have had an affect upon accepting and rejecting the null hypothesis. If confidence intervals had been available in each of those cases, he could have more easily made comparisons. If you know in this study we have a confidence interval of this much, I think it facilitates comparisons, rather than having a dichotomous decision of accepting or rejecting the Ho/null hypothesis.

Mann: Okay. Could we just try to touch on attenuation as an issue over here? With certain exceptionalities, you do have bunching up in the lower end of the score range. In physical education research, which I am reviewing, there is repeatedly to be the claim that there is very little relationship—correlation—between intelligence and motor skills in the retarded, based upon small samples from defective populations when you do have this bunching up.

Hopkins: I am not clear on the question. Are we thinking about correction for attenuation or attenuation in a non-statistical sense there?

Millman: There are occasions when you want to study the relationship between two variables on a restrictive group like mentally retarded. That is, their scores are restricted to a narrow range and that this speical nature of the group may reduce the correlation between the two variables, and to make matters worse, the instrument may be unusually unreliable, which may reduce the correlation. And how do you deal with these problems? Well, there are almost two factors that I see here—which are not independent, by the way. The correction for attenuation is a procedure which allows you to estimate what the correlation would be between two variables that you are interested in, if the two variables had a different reliability than what they presently do. We know that when there is a measurement, it has an error component in it that serves to reduce the correlation. So, people are sometimes interested in what would happen, had we used more reliable measures so we can estimate, if you will, the true functional relationship between these two variables. Most frequently, what is done, is to use the correction formula which estimate what the correlation would be if the test were perfectly reliable or, if, for example, you had an infinite number of trials or an infinitely long test. One has to interpret these resulting correlations—which, by the way, are higher than the original correlations—with some care. I doesn't necessarily mean that you can predict one from the other, at that degree of relationship. It merely offers, perhaps, what is the maximum amount of, if you will, functional relation between these two traits. I think it is a good procedure and one that one ought to follow when one is interested for example, in knowing how much trait A would correlate with trait B if, and this is important, if we have a perfectly reliable indicator for both A and B. It helps you to separate out the influences of two compounding factors. Whenever you get a low relationship, you don't know whether it is because the instruments were not very reliable, or whether there truly isn't much of a functional relationship between these two variables. By correcting for attenuation, you beg in to get a fix as to how much of this low relationship is due to the unreliability, and how much is due to the fact that the facts of the case are that these two traits are very much related. Well, that is the unreliability. However, even if you had perfectly good measures and you are not particularly concerned about the unreliability, you are kind of concerned about the fact that you got a restricted range. You could... there are, again, formulas which make certain functions, or, now, what would happen...what would be the relationship between A and B if we didn't just have a mentally retarded group, but we had people representing the complete range of scores on both traits A and B? I am not sure that is going to be of too much interest. But, clearly, variables just don't correlate as highly when you have restricted groups. I If I had a gifted person and an average person and a mentally retarded person in my samples, and I correlated their IQ scores with most anything, I would probably get a perfect correlation. But, if I had 3 people in my sample having, say, IQ's of 70, 71, and 72—a very narrow range—my correlations, of course, would be very low. There is no answer to that except to get a sample with a wider range.

Feldhusen: I guess the question is sort of like, in a way, related to the one I raised before. This extreme restriction of range, here, we might accept as representing real homogeniety of the sample, and we may do so inappropriately or incorrectly. That homogeniety may simply be an artifact of the insensitivity of the instrument, and so once again, the same rule comes out, as far as I am concerned. I would look if it were possible for another instrument, or further work on that instrument, or some other measurement procedure, which might better represent the sample or characteristics in which I am interested, rather than just blithely assuming and hoping to patch up on the basis of a poor and insensitive measurement procedure.

French: I think we would want to check and see how sensitive the measuring instrument is, and that emphasis ought to be on validation studies with the instruments that are available and the samples we are most concerned about, instead of trying to get into the estimations.

Hopkins: Well, I would just like to say that I agree with these comments. Just to amplify on a couple of points: I don't think it makes much sense to correct for range restrictions. I do think agreeing with Jay, that the correction for attenuation can be useful to get a feel for the theoretical relationship. It is not of any practical value. When one is exploring the real relationship between two variables, I think, he wants to know: What is the relationship that I can get my hands on? In addition, what is the relationship among preschoolers on these variables? I think the correction attenuation can be helpful here. I think there are some real hazards; however, first, the method by continuation.

# CHAPTER 18

# MEASURING CHANGE

Overview: The topic of measuring change is particularly touchy in methodological circles. Two of the most common methods are (a) repeated-measures analysis of variance with blocking possibilities (in the presence of randomization), and (b) covariation of pretest scores upon post-test scores (when initial differences among treatment groups are highly likely). There are a host of variations upon each of these two techniques, consisting mainly of complicated weighting and prediction formulae. Then, there are those methodological cynics who believe that it is simply impossible to measure change appropriately with the present machinery. This chapter considers the more general issues. First, it seems agreed that if a priori different groups are involved in a study, then the estimates of true difference scores are best approximated on the basis of treating the groups separately rather than combining the disparate groups. Second, if one is interested in group averages for purposes of comparing different treatments, then it is probably most appropriate to use raw pre-test and post-test scores in the framework of repeated-measures analysis of variance. Third, unadjusted data is preferable to adjusted data so that wherever covariance is probably to be deferred to repeated-measures analysis of variance. Fourth, when there are inherently different subgroups within a study, there are possible scaling problems on the criterion test, with the low groups at one end of the scale distribution and the high groups at the other end. Finally, wherever matching of some type has been employed, estimation of the true change scores can be affected by regression (see Chapter 20, however, for a fuller treatment of the regression phenomenon).

Mann: Alright. Let me give you an example of the type of problem we meet, in trying to measure change in special education. Zigler says we have essentially two types of retarded populations. One consists of individuals who have a mildly normal variation—they are slower than the remainder of the general —but they are part of the normal curve of ability for that group. Then you have a genuinely defective population. This is a separate population. Its members are members of a distinct group, and not legitimately part of the tail of the normal distribution.

Stanley: A genetic, congenital-by-defective, and so forth, population.

Mann: Down's Syndrome, which you can define as being unique in certainly many characteristics...now, in terms of establishing change, what would you use for the one educable group that is part of the normal distribution of abilities with 70-80 IQs or whatever, and what would you use for this Down's Syndrome? A simplistic, concrete question, no doubt.

Stanley: But a very good one. Well, clearly, the Down Syndrome is a priori defined as being different, identified at birth, and so forth, as looking different in ways that predict lower status. My intuition would tell me simply to treat the Down Syndrome people separately, estimate the true scores quite separately from the others—have two populations—work with them quite separately. Any time you can define a population a priori; I'd do the same thing with blacks versus whites, particularly in what we know about the difference in their means.

Millman: Isn't there a question about that?

Stanley: Well, suppose you are going to do an experiment, and over here you have a group of people—Down Syndrome—and over here you have another—a non-Down group—you can't randomly assign to the experimental treatment. This is one, like having cottage or class now, and this is another. You are going to treat this cottage one way, and this cottage another, because of their differing characteristics. Then, pretty obviously, you cannot contrast Down people in any way with non-Down Syndrome people. You are just using intact groups.

Mann: What would you use in that case?

Stanley: For achieved scores? I'd go back to Lord's and Cronbach's and Furby's article, and use the progression techniques for estimating gains.

Mann: Now, how's about comparing the Down's Syndrome to a normal group in developmental studies, conservation, or whatever.

Stanley: What are you trying to ask, there? Whether the Down's people gain more over a period of time than the normals gain over a period of time?

Mann: Let's say, how could they gain conservation? An experiment with Down's people with MAs of 5 compared with normal children of age 5, and, we are studying conservation.

Stanley: At what age do they gain it, maybe.

Mann: We are dealing with the mental age.

Stanley: Maybe you just have a question of the percent of conservation gained by 5 and you just have an ordinary proportion problem there. You don't seem to have a change problem there, just a proportion—if by mental age 5, 70% of normals and 10% of Down's, you can easily test that.

Mann: Let's say, I want to see how I can improve their locomotion or their balance, by having them walk on a balance beam of some sort. So, I put the Down's Syndrome children, and the normals, through this treatment, and I compare them on this beam again. Now, how would you handle the change, as for the gains?

Millman: Both groups get the same treatment?

Mann: Yes.

Millman: You know how well they could walk before, and how well they can walk afterward. Then the question is, how do we compare their post-test scores—or, how do we compare the effect of training of these two groups in the fairest way?

Stanley: Do an analysis of variance: raw scores and raw gains, and so forth. You have a whole group in each case. You don't have any regression; the average regression effect will be 0.

Millman: I'd say there would be no regression.

Stanley: I'd just say, classify them as mongoloids versus non-mongoloids, or Down's versus non-Down's, crossed with pretest and post-test repeated-measurements nested design, and just do an analysis variance. You are not trying to make a statement about an individual, other than by means. That would be a fairly powerful design, too. Particularly if you had equal numbers of Down's and non-Down's.

Millman: Would you object to doing an analysis of covariance with the pretest scores as covariance?

Stanley: You could, but you could get into Lord's types of problems with that. You see, the beauty of the non-covariance procedures is—if it has any beauty at all—is that it doesn't fiddle with the data or anything. It is a priori—you are saying one times this minus one times this, and that may get around a lot of problems. One could also work with differences and see if the mean difference of one group is different from the other group.

Hopkins: I think this illustrates the issues raised by Lord's. If we just use gain—raw gain scores—and the normals start off at a much different place than the retardates, then we have a lot more on the scaling problems, potentially, right.

Stanley: Except that we are dealing only with the differences. So, in the formal analysis of variance situation, all we would be concerned with, at the most, is whether we have more heterogeneous variance in one group than the other. And, if the groups are of about equal size, we have got a lot of power for that problem, and need to worry quite as much. We do have to worry about the nature of the interval scale—whether you have an interval scale that bridges both of the groups.

Millman: The hypothesis is that the hypothesis is without an interaction. An interaction could occur by scale artifacts, however.

Stanley: Right, that's right. The test is an interaction test. It is the interaction of the classification of pretest, post-test, or the difference between the differences. Now if it's there, of course, you can try to remove it by transforming the scale. You can see whether you can, using the Cox and Box type of argument, see whether you can remove the interaction by transforming the dependent variable.

Millman: The very hypothesis you are trying to test.

Stanley: Okay, that may be, but if it is the question of a scaling problem, it is not something immutable. If Down's children gain more from pretest to post-test than the others, that is really what you are really interested in, isn't it, Jay? There is a question, as Ken suggested, at the level the Down's are. Do the 5 points there have the same meaning as 5 points at the higher levels? That is the question of whether we have an integral scale or not. Now, it is quite reasonable to suppose that for some measures, this isn't true; for instance, if we have floor effects, or ceiling effects. We have got to have both of these groups—Down's and non-Down's—at least poised in the middle of the scale—not at the floor, not at the ceiling—so there is no bouncing on the floor or against the ceiling, and that could be looked at in itself. My guess is that if you have a wide enough range of test, so you know there is no floor or no ceiling, that you minimize the probability of a non-integral scale. If you have some nice thing like height, take height. Do

Down's Syndrome youngsters gain as much in three years as normal children? Well, we know height is more than an integral scale. It is a ratio of scale, too. I don't think we have to worry. We have plenty of floor and plenty of ceiling with height measures, so we wouldn't have to worry. The scale is clearly okay. When we are dealing with something like IQ, where the Down's youngsters are down at non-verbal items that are at a lower level, and the normal youngsters are up in the average adult level of the Binet, we may worry more about whether there is an integral scale that cuts across all that distance. Maybe there is an integral scale that is within the Down's groups and maybe an integral scale within the normals, but it may not be the same type integral scale. There may be some discontinuity there.

Mann: Do you have anything else you wish to add?

Stanley: One of the best discussions of this little point is, simply, in Jetson's famous-infamous 123-page article. In part of it he discusses IQ and integral scales and gives some procedures to resolve the issues at it. But, he doesn't deal with two population problems, which is what you have there. What we have said, Les, what we have said is that you have got to be cautious and look at a lot of things as you go along. You can't just accept on face, you know, you have to worry about the scale and you have to look at the ceiling and floor effects, and a number of things like that.

Millman: I'm not sure I'm satisfied. A practical question that I would like an answer to is when you do matching, do you always have to worry about the regression effect, or can you sometimes match groups-use covariance analysis and feel a little more comfortable when you go to bed at night. I am trying to get clear about situations where you can and can't.

Stanley: Well, let us take the popular issue of the day. Suppose we do mix the Whites, Blacks and Puerto Ricans and everybody together, and we have one big population-and we call it a population. Okay. We test that population, and we work out the estimated true scores on the basis of the reliability coefficient for all that population, the single population. This population can be spoken of in terms of subgroups. We can always identify subgroups of redheads, blue eyes and so forth. But, some kind of experiment in which we are going to match subjects from this population is they must have the same obtained score to be matched. Alright, in that case, we are forced to ignore all subpopulations. We are not allowed to match blacks and white, for example, because we have not used blacks and whites in defining the population. We must match two people if they have IQs of 85-one black and one white, or both black we use them. One gets this and one gets that. There is some kind of judgmental thing required for the little experiment-judgmental thing to decide which one of those two 85s get one thing and the other could be randomly assigned. No problem. Okay. So, that way we are different, Jay. We are not saying, we are going to match blacks and whites and then treat them as one population; we can't do that. We are going to say we have a population, and we are matching any two people on basis of the same obtained score, and their regression will be the same. Is that what you want me to say, Jay?

Mann: Let us consider the EMR subgroup which is part of this normal distribution, according to Zigler. Anything else to be said about this?

Millman: Point of information. Now, I am not going to make a speech—just try to understand better the question. We are in the situation where you want to compare a group of mentally retarded youngsters of this type with a group of normals, and do you propose to do any matching at all?

Mann: Example: you have a group of EMR youngsters—they are drawn randomly from an EMR population—we have normal kids, drawn randomly from normal groups. Question: Are the groups comparable, in terms of response to motor training of some sort, or whatever? In some literature it is suggested that there is no difference in response to motor training in higher level educables. So, now we compare these drawn random samples, if you will, from representative populations.

Millman: Do they differ on some sort of test scores?

Mann: Well, let's say that we have chosen them on the basis of the same chronological age, but, I think, are obviously different.

Millman: Chronological age. Well, they would be matched on chronological age—the two groups. And now, you test them later, and you want to see if they would differ on some sort of test measures.

Stanley: Exactly. You have two different intact groups and you want to do an experiment and compare them?

Mann: The hypothesis is that, based on some research findings, that the response of the higher educables to physical training is really the same as that of normal children. They should make the same gains in motor skill development. So, I compare a certain random sample of 14-year-old EMRs with a random sample of 14-year-old normal children. So, how will I manage this, in terms of the design, in terms of the best yield?

Stanley: You want gain, or change, scores? How does that differ from the problem we had a minute ago? It seems to me to be essentially the same problem.

Proger: I see you are seeing this as separate populations. Okay, but as I suggested, they might be considered appropriately as being samples from the same populations.

Stanley: If you define them as separate populations, sample them separately.

Millman: That is going to change your problem very slightly, and I think, present a different problem than Julian appears to see! That is, let's suppose you give all your—both your samples from both groups—a pretest on this motor skill and your mentally retarded group, just as a group, is not as good as the normal. So you say, I want to give some kind of training. I want to take, for purposes of my experiment, matched pairs from both groups. In other words, I am not going to match them on chronological age at all, fifth-graders at all; they are all the same chronological age, or something. I want to match them any way before I do some kind of training, so that I can say that my sample of mentally retarded was at least initially as good, on observed score of this motor test, as my other group. Now, I want to train them and see if they can both learn as easily well—improve equally as well on this motor skill. Now, how about that situation?

Stanley: Now, if they are very different, that moves your reliability coefficient way up and causes regression to be less. But again, we'd get back to the problem if we can't treat them together, then treat them separately!

Millman: But we have picked the better of the mentally retarded to get the training, and we pick the poorer of the normals to get the training—and they now get training. They now get post-test scores. Now the question is, are the post-test scores directly comparable because the kids were matched on the pre-test scores, or is there a bogus difference expected because each group has now regressed toward their own mean? Is the post-test difference not directly interpretable, because of the regression artifact?

Hopkins: I think you have a regression artifact, and capitalize on measurement error, since the population—your unbiased estimates of population means—are different.

Millman: Ken, I'd say not.

Stanley: There are two different ways to estimate the true scores of these people, and they don't yield consistent results. Okay. One way is to treat the whole group as a group and have a bigger reliability coefficient and a relatively small regression. The other is to treat the two as separate populations a priori before you test them, and then use information to regress each of them separately, in which case there would be a lot more regression for each because the variability would be less for each, and the set of true scores for one will not agree with the other. Now, the question is, which is right? In a sense, a person has presumably only one true score on a particular test. A true score is something he carries around with him. It is not carried around by other people. Now, what I think Ken is saying—what I kind of feel here—is that the true score that characterizes him will be better estimated by the within subgroup estimate, than it will be the other.

Millman: Unless you argue something is happening to these mentally retarded people!

Stanley: You have a very subtle distinction, Jay! We will agree that if we define the population any way we want to—if we define it as a heterogeneous population—then we will allow people, at random, with the same score to be chosen without regard to subgroup characteristics, and then put an experiment under control; I will then agree with you. But, I won't agree that you may then exercise a subgroup choice, too. Because if you exercise a subgroup choice, you have got, I believe, yourself in this other bind. Okay? I agree with you as long as you ignore the other characteristics—any other characteristics—because after all, in defining the population as heterogeneous, you have de facto ignored subgroups' characteristics.

Millman: I don't quite go along.

Stanley: What you are talking about now, Jay, you are now talking about picking a type A and a type B with the same score.

Millman: Type A and Type B?

Stanley: Because you have chosen them for, you know, because they belong to different populations. If you have different means, or assume different means, you have different populations. If they are a single population, you are not allowed, by the logical thing, to look at them separately.

Millman: Seems like we are spinning around.

Stanley: We keep shifting back and forth. We'll start out and we'll say we'll have an undifferentiated population—any two people chosen at random who have the same score, have the same expected true score. Okay. So far, alright. Then we shift, and say, now let's differentiate the population and let's look at some subgroups which are quite different, with respect to average score, and let's match them. I think, at that point, you have lost your gain, probably.

Mann: May I ask: Why don't we match these individuals again, these EMRs, on another variable than the variable we are interested in assessing the gain scores on? We matched the EMR group with the normal group on mental age on a standardized intelligence test, because this is a standard procedure, done very often.

Stanley: I wrote an article criticizing this.

Mann: I know you did. It is still being done, so it's worth considering.

Stanley: Are you measuring matching on some concurrent type thing, or something antecedent to something? That seems to be the problem. It is the old paradigm of the 14-year-old brain damaged boy versus non-brain damaged. Do they differ in arithmetic reasoning, and if so, do they differ because of the brain damage? And, we are going to covary out, or match out something that is fourteen years old. You see, we are, in effect, matching out the brain damage or something. There's a problem here. If you match them before they were born, or something like that, maybe that would be alright, but not fourteen years after the fact of brain damage.

Mann: How about a group of slow learners? We have matched them on this IQ variable, or MA variable, in AM division. We shouldn't match them in the first place.

Stanley: No, I am not sure of this, but I don't know what question you are asking me. It's clear that if we have that simple difference situation we talked about, then we know what we are asking. Do Down's Syndrome kids grow—gain more height in three years than normal kids do, in a certain three-year period? That's okay. No matching problem there. We are just asking a question about differences. We could ask it. Now, on the other hand, if we match them on height initially, what are we asking? If the Down's Syndrome youngsters were as tall as normals, would they...?

Millman: Lee was talking about mental age. Let me try what I think. The question might be: When you can match a mental age, you are, in a sense, trying to equate the two groups on their total amount of, not brightness, but on the total amount of what they may have learned—mental age, you know—how much they learned. Now, the kids are going to be of unequal age, of course. Now the question to be asked is: Are there, in terms of this other thing, are the two going to be equal, now? I saw one study that investigated moral concepts. Maybe the research asks: Is your sophistication about a certain kind of moral concept really dependent upon how much experience you have—your mental age—or, does it depend on something like brightness, or something like that? To look at that, you take these two groups who have quite different brightness but are equated on mental age, to see whether there is a difference in their moral developmental stage, or something like that.

Stanley: That's the old Coldstor and Blysma type of thing, back in 1951. There's some old work done in Iowa, under Stroud, way back yonder. It's difficult to know. These are really important questions. One, we are asking about matching. One just has to worry about the regression problems when we match. If it is height you are matching on, there's not much worry. If it is mental age, more worry, and something else more, depending largely on the reliability.

Hopkins: I agree.

Mann: Now, how about in terms of gaining from treatment?

Millman: I'd say the...I would certainly agree, the regression effect when it is a source of bias, is more worry when you've got unreliable data like MA than when you've got reliable experience like height. I would agree to that, but whether you have to worry about the regression effect of these two populations when you are matching—I would say, no.

Hopkins: Mental age is essentially raw score, isn't it? We have the same regression effects in our subpopulations, don't we?

Millman: If you treat them as subpopulations having different regressions.

Mann: I guess that resolution of this issue isn't in the cards. Enough!

## REFERENCES

Feldman, K. A. Some theoretical approaches to the study of change and stability of college students. Review of Educational Research, 1972, 42, 1-26.

O'Connor, E. F., Jr. Extending classical test theory to the measurement of change. Review of Educational Research, 1972, 42, 73-97.

Poor, D. D. S. Analysis of variance for repeated measures designs: Two approaches. Psychological Bulletin, 1973, 80, 204-209.

# CHAPTER 19

# STATISTICS: PECULIAR PROBLEMS CAUSED BY THE NATURE OF SPECIAL EDUCATION RESEARCH

Overview: This chapter covers some miscellaneous problems in statistics caused by idiosyncrasies of special education research. Three topics are considered: (a) disproportionate cell frequencies in factorial analysis of variance, (b) Bayesian statistics, and (c) factor analysis. First, some pointers are given on how to cope with unequal cell frequencies in analysis of variance. When one attempts to control by blocking on one or more variables (such as IQ or sex) in comparing two or more treatments, the typically small, total number of subjects available usually means that an uneven distribution of subjects throughout the design data matrix will occur. Analytically, a problem arises in such a situation in that the between and within sums of squares do not add up to the total sum of squares. Available computer programs for disproportionate cell frequencies are discussed. If unequal frequencies have been caused by attrition from an originally equal -N matrix, then methods of data substitution are a possibility. Second, the topic of Bayesian statistics is discussed briefly, but the opinion is offered that classical statistics will remain predominant until "cookbooks" can be devised for this relatively new area. Third, factor analysis is brought up, not as a means of descriptive statistics, but as a mechanical device of massive data reduction. Precautions are raised about the ratio of variables to number of subjects, something that special education researchers must be particularly concerned with.

Mann: Everyone is recovered from our last considerations. Shall we go on to unequal Ns?

Proger: In a lot of special education studies, there appears to be a deliberate avoidance of building in blocking variables. Let's say we have EMRs and normals. We want to compare them for some reason or another. And we run into the problem of small and unequal Ns. Let's say we build in a previous blocking variable of reading ability, or something like that. Suppose we have so few bodies in the total study that we have to use almost all of them. As a result, we now have unequal Ns. What procedures do you recommend? What is the best method of treating unequal Ns in, let's say, analysis of variance?

Stanley: General Linear Hypothesis model—Bock and Finn's work. There are any number of computer programs that generally can handle the problem.

Proger: Like unweighted means analyses?

Stanley: The linear hypothesis is more general. They'll manage the whole thing for you in a more sophisticated way than Weiner's simple unweighted-means analysis. So, if you want to conserve the little data that you have, Jay, why not just go to a standard program?

Proger: How about missing data then? That is another issue.

Feldhusen: Why are you in this fix? Perhaps, somehow, I missed some of your explanation. How did you get into this fix on unequal Ns that is troubling you?

Proger: I don't have an explanation, specifically. You could have unequal Ns where you don't have enough bodies to fill up cells. Or you start a study and a class member is institutionalized.

Feldhusen: That is why I asked the question. I would be concerned over the rapid movement into this easy programming or computer analysis for this attrition problem. You ought not to proceed blithely into this analysis, but to worry about the possible considerable bias in this attrition that has occurred which has to be dealt with in some way or other. I would begin specifically by asking you to assure me that what you have lost by your attrition is not different from what remains if you are going to continue to use that which remains.

Stanley: That is a most important point. I was just assuming that you had some natural situation where somehow the attrition was unavoidable and with no defined bias. I think John's question is important! Why have you got that situation? Usually you have that situation because of some kind of correlated variables situation where you kept something as a result of a scatter plot. First of all, you might not be doing the analysis of variance. You'll probably be doing regression analysis or something a little different. I know that analysis of variance is put in regression form, but sometimes instead of trying to throw out a case, you are really not properly using analysis of variance in the first place. You should be using some other type of more general regression technique that doesn't have this proportional subclass frequency probability. Secondly, suppose all your factors are not factors at all, but are things like socio-economical level, three levels—height, in five levels—you might as well use socio-economical level indices and height indices and do a regression analysis in the first place.

Proger: You think your attrition is unrelated, then, to anything in the treatments, or something like that? I believe Weiner says you could use a crude approximation of unweighted means. Is there anything better, though?

Stanley: There are elaborate programs. You can read Box's 40-page article, but there are programs and computers that can do this. But, I would urge you first to look at the attrition problem, and secondly, to decide whether or not you really need to know about the pattern of loss of data, and thirdly, if attrition is okay and everything else in the design is proper, then consider making proportional subclass frequencies. Because then you have orthogonality and the sums of the squares add up to the total. You have a little better situation. You don't need a special statistical technique if you have orthogonality. If you have proportional subclass frequencies, you could do a general analysis— in particular, a fixed-effect model analysis. There are lots of things. Of course, you can throw out data, sometimes. If you have a lot of data, you can throw some of it out; make things proportional, or equal. The difficulty with disproportionate subclass frequencies designs is that they become vulnerable to the violation of assumptions of analysis of variance ofttimes. The best situation is to have equal subclass frequencies. But, if you have disproportional cell frequencies, then you have certain problems, no matter how you do the analysis. You have certain problems of identifying what really was significant, because it makes a difference which effect you look at first, which second, which third, and so on. If you look at them in reversed order then you may get a somewhat different set of mean squares and _F_ test. If you put sex in first, race second, and socio-economic third, you get one set of mean squares. However, if you reverse it so that you have socio-economic level, race, and sex, you get another set of mean squares in the nonorthogonal analysis. This is true of the Bock and Finn program, because it is a stepwise multiple regression situation. Gerry Finn ran the program for me on a similar problem and had to redo it two or three ways just to see whether he was getting different results. He had to reorder the main effects. Gerry is very clear on it. In my design, Gerry said, which one do you want to estimate more precisely? We will put that one in third. Let's say I want to estimate sex-best. We'll put in college first since I'm least interested in it, then race with college held constant, and finally sex with college and race both held constant. That is basically what we have got. Now in an ordinary analysis variance with equal frequencies among cells, that problem does not occur. In the latter orthogonal case, it would not make a difference. You could use them in any way you would like.

Millman: But the Bock-Finn approach is not the usual analysis of multiple regression analog analysis of variance in disproportionate cases. That is not what Ward Bottenberg does in his technique.

Hopkins: I think it is a misnomer to talk about the Ward Bottenberg approach. In isn't his approach; they developed a computer program. However, a mathematical statistical redeveloped the ration, but it is multiple regression.

Stanley: Janet Elashoff out at Stanford and Alan Abrams worked up an alternative procedure which does the same kind of thing. I haven't looked at it or used it. My impression was that intrinsic to the nonorthogonal situation is this indeterminancy.

Bracht: I am uncomfortable with the multiple regression approach in the analysis of variance design where there is a disproportionate _N_ becuase, at least on the interaction effects, I'd seen some illustrations where you can get an interaction effect even though the means themselves do not show any interaction. I still feel uncomfortable with disproportionate Ns.

Stanley: Now wait a minute. I'm trying to think here. In the old days, before computers and so forth were very common—this is what you get for living so long, you know, you remember the old days—I remember in Chicago, in 1955, people were rushing in from the Psych department, from the Statistics department, with negative interaction terms because they had disproportionate frequency designs. They were doing ordinary analyses, and as they did in those days, they were getting their interactions by subtraction from the total sum of squares. Things sometimes turned out to be quite negative. This, obviously, is the root of the problem. Now, how much you get away from that with these special programs is something that I have not been quite clear about. This nonaddivity means there are some funny cross product type terms that are getting clustered in there. You know, like sum of squares A times sum of squares B type terms.

Proger: We were talking about predictions. Let's suppose we are trying to derive a prediction equation on a group of EMRs and we have some trouble getting complete data. Suppose we have many predictors and we do a pretty beautiful job of gathering the data, but we wind up with one or two of the predictors missing on a few of the people—what do we do? Drop the people, or...

Stanley: Marian Shakoff published a paper on that technique. She shows you how to condition a matrix wherein you base your correlations on different numbers of errors. Then, of course, you have the same problem John mentioned a minute ago. Why are those people missing? Are they atypical people?

Hopkins: I think maybe what we are saying is not that realistic. I think whenever you try to collect data on a large number of variables there are going to be gaps in it, and you can either try to estimate what the score would have been, or you can just base the correlations on those sets of data you do have, and recognize their limitations.

Stanley: Of course, the simple way that so many people use to solve this problem is to eliminate everybody who didn't have a perfect set of data. But you have a biased, restricted, and much less generalizable group to work from, less variables and so forth. Well, I think that is undesirable, particularly if just an observation or two is missing from the group. If one person in your data has a lot of observations missing, then there is not much good you can probably do for him, but if he has just one observation missing, you could complete multiple regression approach to predice that one observation among all the other data that you have. That would be—in the old days—unheard of, but maybe nowadays with computers there exist programs that would do that for you.

Hopkins: You may remember the Allenson data we collected. We defined another variable as being simply the presence or absence of complete data on any given person and then correlated that dichotomy with all the other variables to get a fix on whether or not there may have been certain kinds of selectivity.

Feldhusen: This is an appropriate time to echo the earlier assertion to watch your data, also. If someone else is doing the recording and coding for you, watch out! Lo and behold, it seems most tempting for people who are recording data, or finding missing data, to put zeros in the boxes. Only after the fact, in several cases, after much searching, have we found that those artificial zeros were the cause of distorted results.

Mann: There are a number of extra issues, here, that some people's tests have raised. One is Bayesian statistics.

Stanley: It is like nonparametric statistics, it seems to me. A certain contribution has been made, but I haven't seen any tendency for it to take over. The Bayesians are certainly strong advocates of doing nothing else, but I don't think our textbooks are reflecting it yet very much, are they? In spite of Don Myer's fine work at the various AERA presessions, I don't know of any of the new statistics books reflecting substantial Bayesian approaches.

Millman: Well, there have been some.

Stanley: I mean in my own applied area.

Millman: It certainly isn't going to overtake classical statistics until there are cookbooks. However, in any massive use of statistics there are judgments of almost Bayesian type that one makes and they sneak in at least implicitly, if not explicitly. The

Bayesian approach sort of makes the judgments more explicit, and makes them in the beginning of a study. I think that the traditional approach often fails to recognize that you need to put your decision-making in the beginning. I think of the excellent example—what I thought was the one of the two outstanding papers of the six hundred I looked at as the JEM editor—of Dick Darlington's paper on: How Do You Best Predict the Criterion That You Can't Observe. I thought it was a brilliant paper. I still do. And it's a Bayesian solution. There, it is necessary to put your judgment early in the game, rather than late, and I think there are cases where that's appropriate. It may be, however, that the application in special education may come when you are comgining a number of small case studies and you can make use of the prior histories of the subjects. That is one thing that I think the approach is best at—in using prior information in evaluating a present set of data. The approach may be useful in combining studies on similar populations of children of a type who would normally never occur in large numbers in any given study.

Stanley: But you know we have been doing this for many years. Old Donald Petterson (who is now dead, I guess) up at Minnesota, applied psychologist who founded the *Journal of Applied Psychology*, had what he called antecedent expectations. This is a very helpful concept. It's a basically intuitive Bayesian concept. For instance, when I was out of graduate school only a short time, I once computed a correlation between freshmen grades and some test scores and got .86. My, I was happy! I had discovered something that was big. However, my antecedent expectation was lacking. I didn't have the skepticism that should go with it. I didn't have a nice interval in my mind that would suggest to me that outside that interval the probability is extremely low. At any rate, I thought about it a little bit, and everybody checked my work, and it was simply incorrectly computed! It was more like .43 or something. But this is illustrated by old Walter Debbon's 1906 correlation. The first one, I think ever correlated between high school grades and college grades. It came out to be .80. Now nobody had any evidence, so they said, "Oh boy! That's good!" Years went by. People began to report more and more on this topic. None of them were .80. They were much lower than that (.4, .5, .6 at most). Finally, in 1927, twenty-one years lattr, Debbon put out his data which he still had, andhad Kekotell (who was a graduate student of his at that time) recompute it, and it came out to be .60. By that time the overwhelming evidence had caught up with it. I think there are a lot of ways we do this type of thing without having appropriate antecedent or Bayes-like expectations. The question, I guess, is really to what extent such expectations can be mathematized and formalized in inferential reasong. The Bayesian field has been growing pretty good awhile. We have Regershich's book of eight-ten years ago, but it just takes a long time to get into the cookbook stage of psychology, and until it does that, as Jay says, we are not going to use it much.

Mann: Can we take the last ten minutes or so and ask some questions about factor analysis?

Feldhusen: I guess you earlier identified the problem that seems to keep coming up here all the time with factor analysis. And that is the problem of sample size. You speak of just one situation after another in special education where you are struggling to get along with very small samples. I've expressed pessisism why that should be so. I wondered whether it was not undue narrowness in sample seeking, as it were, on the part of researchers. But in any event, if you are inevitably bound to very small samples, then you are quite limited in the amount of factor analysis you can take on.

Millman: One of the most frequent errors, or let's say questionable practices, that I have seen in the manuscripts for various journals, is to do factor analysis on say forty or fifty cases. One I had received was with thirty-six cases. Again, the rule of thumb is that with factor analysis as with any other analytical situation, the believability of the results depends on the initial input. The initial input to a factor analysis is a correlation matrix. The correlation coefficients are not very stable kinds of things when computed on small samples, and certainly a pattern of correlations is even less stable when computed on a small sample. "I've heard some people say that we shouldn't do factor analysis unless you've got some two hundred cases. I'm not sure there is a magic cut-off, but certainly, forty cases or thirty-five cases are far too few. Now, there is another concern in factor analysis, and that is that in deviant populations some of the correlations could be due to the restricted range of talent or ability.

Stanley: Floor effect, Jay, for instance. Some variables have floor effects that are severe, and others will have virtually none.

Millman: That might make a correlation kind of funny looking.

Stanley: Carroll showed years ago - 1945, I think it was, that difficulty factors can show up in factor analysis as factors. Now you generalize that in effect factor coefficients dealt with dichotomous data, but you generalize and say that if the shapes of the distribution of reverse () variables are radically different from each other, you may get artifacts. Factors that artifacts of the fact that brandesque of the variable, can't correlate highly with brandesque positive variable, and that would give you a low correlation. It's important to remember that the correlations in factor analysis can be a function, not of some intrinsic underlying resemblance between variables, but of the way they are scaled. One happens to be a low ceiling, another a low floor, that would give you the point. I guess it is generally true that not only with disadvantaged deviant groups, but with any group, you have got to keep your people along variables where they score towards the middle and not give them plenty of room at each end to expand on.

Feldhusen: Certainly factor analysis offers help in the classic situation where one looks to make a meaningful reduction of his data to a smaller number of variables, hopefully, which enters again is that if the reduction process does work out sensibly, and one has to choose factors that seem to have some substantial amount of variance associated with them, one still has considerable difficulty trying to figure out sometimes what they mean, or what to call them, or what the construct is, or what have you. A very beautiful illustration I invite you to look at is Curetion's study just completed for American Institutes for Research. She started with 700 project Talent variables and very neatly was able to reduce that not to one set of factors, but to blocks of factors in different domains that become manageable and yielded very good interpretable results. So, if N permits, this certainly seems an important solution. I know by experience the horrible problems of trying to grapple with several hundred variables, to make sense out of them, and certainly some other way must be sought, rather than trying to deal indiscriminately with them as is often done.

Millman: There is one empirical study which looked at the effect on factor analytic solutions when you do include variables which have floor and ceiling effects, and you impose transformations on the distributions, but don't change the rank order. This study has shown that factor solutions are virtually unchanged. So that may not be a problem.

Stanley: Is that Belgado's big study? Edward Belgado?

Millman: No, this is the fellow from Marquette, Quarishi, he has one very immense study.

Stanley: Belgado found out about the same thing, but he never published it. These dichotomous-scored variables present a problem. You see the dilemma when 5% of the people get the item right and 95% get it wrong. When this item is correlated with another item that 5% get wrong and 95% get right, the most you can expect is .056. On the other hand, on the item that 95% get right and 95% get wrong, you can get a correlation on one with another item that 95% get right and 95% get wrong. This situation probably makes less difference with ordinary studies that have small numbers of children because the shape of the distribution is a random function of the sampling, as much as anything else. However, there is one commonly used approach that particularly bothers me, not so much in factor analysis, but more so in multiple regression. One of my colleagues is prone to use forty people and seventeen variables with multiple regression equations. Obviously, the overfitting and the blowing validation with an independent group would show you a great deal of shrinkage in the multiple and would bring you back to reality.

Hopkins: I agree with John. The primary use I see of a factor analysis is data reduction, not in describing the way things are out there. I think factor analysis is just fine, as long as you are trying to reduce variables and combine them, and so forth. But in terms of describing the structure of this or that, I don't have much hope that it will ever be able to do that.

Stanley: That is a good point.... subsidary point, here. Do any of you know of studies in which various variables have been normalized before the correlation procedure has been used? One gets rid of the problem where some variables have different distribution shapes. I guess this is what Quarishi was doing.

Millman: Also, we might mention Brown and Peter Dedusort who looked at the distribution problem and normalization.

Stanley: Yes, he allowed some to be skewed one way and some another, and so forth.

Mann: What happened to these differences?

Stanley: Not much.

Millman: But again, it is not the crucial test because he did not instruct some subjects to do the Q sorting with the forced-normal distribution and other group to do it just any old way.

## REFERENCES

Dodd, D.H., & Schultz, R.F.,Jr. Computational procedures for estimating magnitude of effect for some analysis of variance designs. Psychological Bulletin, 1973, 79, 391-395.

Fischbach, T.J., & Walberg, H.T. Weighted and unweighted means for estimation: A note on the Humphreys-Dochler and Jensen papers. Journal of Educational Psychology, 1971, 62, 79-80.

Gocka, E.F. Regression analysis of proportional cell data. Psychological Bulletin, 1973, 80, 25-27.

Light, R.J. Issues in the analysis of qualitative data. In R.M.W. Travers (Ed.) Second handbook of research on teaching. Chicago: Rand McNally, 1973. Pp. 318-381.

Meyer, D.L., McSweeney, M., Marascuilo, L.A., Collier, R.O.,Jr., Larson, R.C., & Tatsuoka, M.M. Statistics. Review of Educational Research, 1969, 39, 723-747.

Novick, M.R., & Jackson, P.H. Bayesian guidance technology. Review of Educational Research, 40, 459-494.

Overall, J.E., & Spiegel, D.K. Comment on "Regression analysis of proportional cell data." Psychological Bulletin, 1973, 80, 28-30.

Perlmutter, J., & Myers, J.L. Comparison of two procedures for testing multiple contrasts. Psychological Bulletin, 1973, 79, 181-184.

Proger, B.B., Taylor, R.G., Jr., Green, P.A., Cross, L.H., & Mann, L. Unequal cell frequencies in analysis of variance: A review and extension of methodology for multiple missing observations. Paper presented at annual meeting of the American Educational Research Association, New Orleans, 1973. Available in ERIC document ED-074-085.

Pruzek, R.M. Methods and problems in the analysis of multivariate data. Review of Educational Research, 1971, 41, 163-190.

Smith, I.L. Comment on Joe's "Comment on Overall and Spiegel's Least Squares analysis of experimental data. Psychological Bulletin, 1973, 79, 170-171.

Tatsuoka, M.M. Multivariate analysis in educational research. In F.N. Kerlinger (Ed.) Review of Research in Education: 1. Itasca, Ill.: F.E. Peacock, 1973. Pp. 273-319.

## References

Tatsuoka, M.M., & Tiedeman, D.V. Statistics as an aspect of scientific method in research on teaching. In N.L. Gage (Ed.) *Handbook of research on teaching*. Chicago: Rand McNally, 1963. Pp. 142-170.

Winters, J.J., Jr., Proposed analyses of floor and ceiling effects. *American Journal of Mental Deficiency*, 1972, 77, 296-300.

# CHAPTER 20

# COVARIANCE AND REGRESSION
# TO THE MEAN

Overview: Regression to the mean is always a confusing phenomenon, but yet it is one that directly bears on the measurement of change and the area of covariance. An example is posed in the measurement of change where one has two intact classes; the classes are of somewhat different abilities but nonetheless presumed to come from a common population. First, one must note that the phenomenon of regression to the mean is possible in a change situation, regardless of whether he analyzes his data by covariance of pretest upon post-test or by simple gain differences. Second, if the two groups are not really that different in nature, then the tendency toward regression probably is not that great; by the same token, the greater the inherent differences between the two groups, the greater the tendency for regression to occur. Apart from the example and the two main points on regression, some useful miscellaneous observations were also offered. First, most people prefer regular analysis of variance with blocking in comparison to covariance. Second, if initial comparability of groups is absent because of lack of experimental control, then covariance is probably the desired procedure for adjusting such initial differences. Third, in cases where one cannot randomize and where one knows intact group differences exist at the start of the study, some would recommend matching in the classical sense on the basis of estimated true scores and then regressing (or covarying) the obtained scores. Fourth, covariance of pretest upon post-test can be likened to residual gain score analysis. Finally, if classical matching has occurred, then regression to the mean is quite likely and covariance cannot get around it.

Mann: Covariance is the standard procedure when the groups have not begun at the same level. Can we proceed from this point?

Hopkins: I would like to make some general comments. I am sure that we would agree that if you have an experimental group and a control group, it is probably better to covary than to use just the change for scores, comparing differences between groups.

Stanley: Let me qualify this. I wrote an article on this in the Journal of Special Education. The analysis of covariance is a nice, beautiful, useful technique for true experiments, where there is no problem of equating groups; that is, where you have assigned your groups at random to the experimental conditions, and to control conditions. In that case, all you are doing is increasing precision of power. You are trying to minimize the within-group variability by using residuals, rather than using regular analysis of variance, when you have initially unequated groups—not randomly assigned—is a peculiar one. You can get funny results with covariance sometimes. Now, the literature is peculiar regarding true experimental efficiency. Let's say, you worry about covariance versus blocking. Cox had an article back in 1957, which didn't consider the power problem but did show that blocking was better than covarying up to a correlation of .6 and about as good at .8. So in effect, from one standpoint, randomized blocking is almost invariably preferable to covarying, because for one thing, it is not so vulnerable to assumptions of linearity and homoscedasticity. True experiments are the only times you can do randomized blocking. The so-called matching which so many people confuse with randomized blocking is not really the same thing at all. Randomized blocking means that one has the possibility of flipping a coin. He has two units and he flips a coin. One gets treatment A and one gets treatment B. But, on the other hand, if you were going to match people because they already exist as treatment A people, and exist as treatment B people, then you can't ever have a randomized block. Here, the matching is subject to the usual biases and regression caused by the fact that if you are comparing initially different groups, you are going to be matching the different portion of one with another, and their regressions are going to be different. So the best thing is to match on estimated true scores or regressed obtained scores.

Hopkins: Let us say that you don't have any alternative of randomized blocks, and you are stuck with initially different groups. If you use just raw gain scores, instead of covariance, on some kind of pre-treatment measure, you get an artificial analysis, because, in effect, you are working with residual gain scores, in a sense, with covariance. Would you agree with that?

**Stanley:** Well, you probably get around some large problems better with the raw scores. Again, it depends on what you are trying to do. If you have a whole group and are just concerned with means, then probably you are not in too terrible shape with just raw mean scores, are you? On the other hand, if you have matched in some way, then covariance doesn't get you out of regression problems. Take the same interest group problem. Say we have a control group and we have an experimental group, both designated for convenience. They have been obtained in the same school, but they are not randomly selected, nor randomly assigned. We now take the differences between initial status and final status. Well, we know we have lots of bias in the individual measures, since the people who are high on the initial status tend to have high positive errors, and they tend to regress downward. Those who are low on initial status tend to regress upward spontaneously. So, in a sense, if we are going to talk about individual gains, we must talk about an estimate of gain which takes this regression problem into account, using one of the Lord methods. Well, the simplest method proposed was simply that you estimated the true score for each person. Lord and Cronbach and Furby described these. You regress the score downward, in terms of one minus the reliability coeffecient, and then you take the difference between regressed scores. However, Lord showed this is not the most powerful way to do it.

You can take more information into account, in a technical way, as Cronbach and Furby described. We are saying that covariance analysis would be better than simple gain analysis because the covariance weighs the relationship properly, rather than using unit coefficients, in effect, one times this. But, that is different from the situation where one estimated true scores for the whole group. There, the thing kind of washes out, in a way. I think what I can say is that using a Lord-Cronbach technique might be the best way of estimating the change score directly by multiple regression procedure; taking all available information into account. I guess, compared to raw gain scores, covariance is somewhat preferable, but you still have the regression problem. There is a regression problem in the individual scores. All of you know very well, from fooling around with algebra, that the reason we have this regression problem is that the obtained scores correlate positively with errors of measurement. It is mathematically necessary that there be this kind of regression, due to errors of measurement, just as in a more general case there is regression due to lack of perfect correlation. But the question that has interested me is: How do you know that the person whose score you are trying to regress really belongs in that group?

Suppose we have some sixth graders taking a wide-range type of test, and some twelfth grader decided to come in and take the test with them. The twelfth grader would be an outrider, in the sense that he would be beyond the distribution, and yet, we would regress him as though he would have been a sixth grader. His true score is a lot higher than the mean score of that group, and yet, we are going to pull him down 20% of the weight of the means, say. Of course, he does increase the reliability of the test by being there.

Proger: Yet, this is typical of handicapped classes. We have a wide variety of abilities in a very small group, and the children often do not "belong" there.

Stanley: Who has some ideas about that? The problem has puzzled me for many years, and I have not quite been able to settle it.

Millman: You have two intact groups that are different in their central tendency, in some control and matching variables. If you then match and take, say, one person from the high end of one group in order to match him with a person from the low end of another group, you then run the risk of bogus difference, due to nothing but the regression phenomenon. However, true as that is, it doesn't necessarily follow, it seems to me, that you have a regression effect, because in that is the assumption that the nature of the differences between the groups has, in a sense, been responsible for the difference on the pretest score. Let me give you a for instance. Let's suppose that the two groups are composed as follows: the children are in a classroom in a school. We have too many kids in one second grade class, so we make two classes. But the way we decide who is going to be in which class is the rank order of the children, in terms of IQ categories. That is, we take the kids in the top two IQ categories and put them in Miss Jones' class, and the kids in the bottom two categories and put them in Miss Smith's class, and put the kids in the middle categories essentially half and half between the two classes. Now, there is nothing in this situation of Miss Jones' and Miss Smith's classes that necessarily produces a regression effect. In other words, looking at it graphically, it is quite possible that the two regression lines predicting pre and post-test scores for the two separate groups really coincide. Here, there is not going to be any bogus difference. This is simply because when you assume that one line is going to be higher for one group than the other, because there is something in the nature of being in one group that gives you an advantage, then you get that kind of thing. I think that would be my answer to your paradox.

Bracht: I would recommend that readers look at Ken's example of how this can occur in the Journal of Special Education. His situation occurred because he drew the picture with one regression line above the other. In his example he had cultural deprivation, I believe, and so the assumption which, I think, is a perfectly legitimate and reasonable one was that being culturally deprived, in itself, carried with it a whole host of things that raised rates and lowered your regression line, relative to this other group. If you are in that kind of a boat, then the regression effect might occur. But, if you are in the situation Jason described, when you are along the same regression line, or you are a twelfth grade person taking a sixth grade test, then, I think, you have a distinctly different situation.

Millman: The only specification that I adhered to was that the pretest and post-test were unchanged. In other words, if the mean IQ of Negroes or Blacks was 85, pre and post, and of Whites, pre and post, was 100, you will never have a common regression line unless the correlation between all those measures were 1.0.

Stanley: In other words, unless you have no error of measurement, you can't have perfect correlation.

Millman: Actually, it is Fred Lord's picture with a different example, for he uses, I think, weights for males and females on changes pre and post. If you covary, you get a significant difference, and it is the same problem earlier. I think the tougher problem is the one Julian mentioned. It is one I never felt really comfortable with. If you go down and get a mongoloid with a 10 or 50, and get a true score for him, I don't think that is going to ever approximate the score he is most apt to obtain on retesting.

Stanley: It is like the outlier problem in statistics. You have this score in the extremes of the distribution. You would like to throw it away, but you'd be accused of bias if you did. So, you have to develop some kind of rationale and get a statistician to work hard at this, to decide if the score really belongs there or not. If it doesn't belong there on some logical grounds, you can throw it away, or fudge it some way. Perhaps you can transform the score. Now, there is one thing that happens that is confusing. If one has a small group with no IQ below 80, except for the score of a mongoloid, it is going to increase the variability of the scores a lot and raise the reliability coefficient; in turn, we have less regression to the mean. On the other hand, suppose there is just one IQ scored 50 and 5 million other more normal scores, then the 50 won't make much difference. But, in principle, putting the mongoloid there causes me trouble in thinking, because I don't want to regress that mongoloid. A mongoloid is, in effect, a random sample from the population of mongoloids, not from a random sample of a population of kids with average IQ of 100.

Millman: Remember, the key point was that the two groups were inherently different between themselves, but not within themselves from pretest to post-test. Granted, this may be an artificial situation, but there is no way you are going to get a common regression line and cut through both of those means. You have two regression lines and not one. So, if you ran an analysis of covariance, of course you would get a significant F. You know, neither would change at all, immediately.

Stanley: Let me pursue this mongoloid example a little more. Suppose we don't say mongoloid; we have a youngster for whom previously gathered data suggests he will never function near the level of the group as a whole. We have a type problem. Suppose we know by previous anecdotal records, observations, etc., that the person doesn't belong to the group. So, if one really has this kind of "prior" information, then one is ill-advised not to use it. The outlier's data, in such a situation, clearly should not be included in the group. We are not saying that we always exclude extreme scores. What we are saying is that you had better have an excellent rationale for excluding such scores, when you attempt to do so.

Hopkins: Don't you think, Julian, one of the undesirable side effects of the computer is that people don't look at their data anymore to detect obvious inconsistencies, such as outliers. I saw a study where there was a .98 correlation between the music supervisors' rating of kids' music ability with their IQ. But, the investigator really didn't know enough. There was one card mispunched, so the IQ was 900 instead of 90, and the rating was 50. That kind of thing can creep in. When analyzing data on the big machine, we really should be much more careful than we usually are.

Stanley: That is beautiful. I understand, and I have heard it said, that in the Rosenthal and Jacobson book, Pygmalion in the Classroom, the computer was working to identify the absentees as having IQ of zero, thus pushing down the means. Some of the classes came out with average IQs of 43. This was caused by half a dozen absentees, each with an IQ of 0. So, there are all kinds of problems of that kind.

Millman: Let me return to our regression example. Two groups of students are unchanged from pre to post-tests, Ken, on a common test. One group is, say 85 on both pre and post-test, and the other group is 100 on both pre and post-test. They, obviously, must be on different regression lines. That is true, but the ad hoc thing is that making the assumption that the pre and post-test scores are going to be the same is tantamount to saying that there was some other influence that was going on. In other words, you are loading the dice if—for example—I have one class that I just rank on my pretest, and I cut them in half and put them in two separate classes. And then, I just match those few people around the middle there. Then, those two classes are not going to get the same pre and post-test scores. And when you say they are going to get the same pre and post-test scores, you have blocked off some situation from the environment that has caused them to be on separate regression lines—you stuck something in there.

Hopkins: I agree with your example exactly. I don't see any conflict. I do not see a contrived example here, at all; indeed, that is what would happen if we took all of the Blacks in school district A, tested them to obtain a mean score, and retested them a year later. Their IQ isn't going to increase appreciably.

Millman: Ah, but I say that it would. It will increase to the common population mean. You think that the Black population is somehow distinct from the White population. That may be true, but then when you make that assumption, you have to worry about the regression assumption.

Stanley: You seem to be thinking that maybe the Blacks might be regressing to 100.

Millman: That's right.

Stanley: Well, why?

Millman: I am not really saying that they would. I don't know enough about Blacks and Whites. They probably would. I think that the Black population is lower for reasons other than, say, genetic problems.

Stanley: This is kind of aside from the question. There is this issue of practice effect.

Hopkins: What you are suggesting, Jay, is that the difference between mean White IQ and Black IQ is less each year which, of course, it isn't.

Stanley: It hasn't changed in fifty years.

Millman: That is because there is something about being Black and White in our society that has influenced those IQs. In my example, I have tried to take a situation in which everybody agrees there is only one population. And, I want to arbitrarily rank these people on a pretest score and then divide them into two groups. I want to ask whether or not the means of the two groups will be the same, assuming that the pre and post-test means for each of those separate groups are the same. I'd say, no.

Stanley: Let's see, Jay. Let's take a very simple case. You take a distribution of scores and simply slash them in the middle at the median. Now, you retest the top people and retest the bottom—there is going to be movement. The upper group is going to move down to the average, and the lower group is going to move up towards the average.

Millman: But I am going to insist that the post-test scores for both of those arbitrarily-divided groups are going to be the same as the pretest scores.

Stanley: I don't think we are saying that those raw post-test scores will be the same as the corresponding pretest scores.

Hopkins: Let's clarify some issues, Jay. In effect, you are defining your groups by their performance on the covariable, that is, the pretest. On the other hand, I was talking about the adjusted means of your high and low groups. There is a common regression line. There is a difference between zero. But that is a very different situation from what you described. When you have defined your population by a variable that isn't a covariate, then we are into a much different ballgame.

Millman: Let's say, I have a distribution of intelligence scores for a bunch of people. Those that are below 70, I am going to arbitrarily call MRs. Those above 70, I am not going to call MRs. Now, I give them both a reading test. I also give them a reading test later. Wave the magic wand over the children. Call it a treatment, or something. What's going to happen to the pre and post-test scores for those two groups? Are they going to be identical? Of course, they are not.

Stanley: That is the meaning of regression.

Hopkins: First of all, if the variable that you selected to classify children into groups is equally correlated with two subsequent tests, then you don't expect their means to change. You have got to manipulate your correlations to make your point.

Millman: Say, we have Miss Jones' class and Miss Smith's class. These are two separate classes. Miss Jones happened to get the brighter group this year and Miss Smith got the slower group.

Stanley: There will be regression of means. The scores will regress, if there is any error of measurement. But consider mongoloids versus normal children. Now, let's make this our example. You wouldn't expect mongoloids to go up on the second test, would you? They would tend to vary randomly, from one test to another insofar as systematic practice effect of some kind occurs. You don't expect mongoloids to regress toward the normal mean or the normals to regress toward the mongoloids' mean.

Millman: On the same test?

Stanley: Yes, but across different testing sessions.

Millman: I guess the point I want to make is that when you give, as an example, a situation in which each group has been pretested and post-tested, the difference of these two groups doesn't change. My original example simply formed two groups from one large group. In contrast, you are sneaking in the assumption that these are inherently different groups. We are on different wavelengths.

Hopkins: I think the example of scores not changing from pretest to post-test, is not that contrived. Such a restriction is certainly legitimate for weight or height. I mean, there is nothing sneaky there. It is a matter of fact that given a group of adults, their average height isn't going to change that much from pretest to post-test. Similarly, I don't think the example with IQ scores is rigged or unrealistic.

Millman: Wait a minute. Take that male and female. You weigh them now, and you re-weigh them two minutes later. Granted, there will probably be no change across time, but there might be a general difference between sexes, in that men weigh more than women. Now, there is no regression effect there, is there? This is somewhat different from the example I was describing.

Stanley: There is no tendency for the men to change systematically from one measurement, or for the women to. They maintain the same distance. We are really talking about changes here, rather than covariances.

Hopkins: That still fits my diagram. There is no regression line that goes through the mean of both people. There are two populations.

Millman: Why should there be? You know you have inherently different populations of men and women.

Mann: Would you say the same thing would hold for the mongoloid versus the normal? They are from the same population—they have hands, eyes—are from the same human species.

Millman: In terms of the pretest scores, no, I don't think that would be the case.

Stanley: Let me ask you a possibly related question. I have often wondered about regression in human beings. What happens when you look at the children of two people with different backgrounds? What do you regress toward? Pretty obviously, your inheritance is a closed system. I don't regress toward some unrelated person's mean. It doesn't seem sensible. You have your own limited supply of genes. If a person with an IQ of 120 in the Darwin family married a person with an IQ of 120 from the Huxley family, their children would have an expectation of going up in IQ. Presumably, even if you have some low-class Darwin and low-class Huxley on your hands, you would expect them to go up toward the means for the general population of Darwins and Huxleys, not down to the general mean of 100. So, we have something slightly analogous to this in other closed systems. We expect the mongoloids to regress toward the mongoloid mean, or we expect the children of mental retardates not to regress to that mean, but towards the IQ of the general population. Half of the heredity is in the two parents; the other half is in the rest of the family line. The whole sum, which goes back to Adam, includes a lot of people.

Millman: Let me try one more example. This is an example of a school that has two classes for mental retardates. They are put into the classes as the normal constraints of administration dictate.

Stanley: Would you let the children be assigned randomly?

Millman: Not quite randomly, but routine assignment to two different classes. It turns out that, as you would expect, one class was a little bit better than the other, in terms of IQ, or something. Now, one class had this special treatment for their mentally retarded people. The teacher in this one class waved a wand. Okay. It is supposed to have this terrific effect. Of course, we know it doesn't. Then, on the basis of pretest scores, we match for kids the best we can. But to do the matching, we have to omit certain children from each of the two classes, because they aren't quite equal. In order to make this a nice, clean comparison, because the two groups are not initially equal, you take five kids out of one class and you match them with five kids out of the other. The five kids you take from one class tend to be below the average of that class, and the five kids from the other class tend to be a bit above average of the first class. Now, you have matched pairs and you do a comparison based on the posttest. Is there any reason to believe that there is going to be any regression?

Stanley: I think so. It seems to me to be a classical regression paradigm.

Hopkins: I think they are both regressing in the same direction. I think the only real reason that the two classes are different originally is because you specified a kind of quasi-random distribution of the total retardate population. It was not really a distinction that in any way defined conceptually separate populations.

Stanley: We have a problem of sampling here. If you form both classes strictly at random, and one class would, in fact, have a higher mean than the other class, this is entirely a sampling phenomenon. Now if you match, you will have the same standard regression problem, as far as I see. With the people who are really high, if you take the lower ones of those, they will move up toward the mean. If you take the upper ones of the lower group, they will move down. Even though they are matched initially, they will lose the matching on the retest, it seems to me. What is it you don't see?

Millman: I think they are on a common regression line.

Stanley: Let's forget the regression line for a minute.

Hopkins: There is such a small difference. There is almost one or two cases that you can find to fit your situation, where one is below his mean and the other is above his mean.

Millman: Would you like it better if I said that everybody with brown eyes is put in another? Now, there are two distinguishable kinds of groups.

Stanley: Now, if you don't have any difference in the means of the groups, then matching won't cause any systematic differences. That is certainly true. If the two groups that are being matched do not differ, then one can line them up, one-to-one, all the way down the line. One has no problem anymore.

Millman: Let's continue this later. I think you and Ken are wrong, in this sense. I think there is an assumption that you are making when you claim there is this regression.

Stanley: I wish we had a blackboard. There is a little algebra that shows the necessity for this whole thing. It is not an empirical observation; it is a statistical necessity. The overlap of obtained scores with errors that they contain make it necessary that, in certain situations, there be regression toward the mean in either direction. If one chooses groups that differ, with respect to the mean on some pretest measure but that have the same variability and so forth, then one is forced, when doing the type of matching Jay described, into a regression situation.

Millman: You keep forgetting that I am talking about children who come from the same population, rather than from inherently different populations.

Stanley: Okay. Even if they come from the same population, suppose you assign five cases at random to one group and likewise to the other group. You've got ten people, and they differ a lot in IQs, say. Alright now, the chances we will get pretty disparate groups are high. There are errors of measurement in the original ten. We have a reliability of say .8, or something like that. There are errors of measurements in the two groups we have set up. Now, those groups will have quite different means, and they will have quite different regressions. They must, it seems to me, regress toward their respective means.

Millman: Out of those ten who have been divided into two groups of five, there are two people in each group who happen to have the same IQ.

Stanley: Okay. Now, the two highest in the low group could be matched with the two lowest in the high group. That is basically what you are saying. Well, the two lowest in the high group will regress upward on retest. The two highest in the low group will regress downward on a retest. They must—I mean, that is the whole object of regression. At least that is my analytic claim. Do you agree with that, Ken?

Hopkins: I won't say yes or no, until I am absolutely sure I have grasped the situation.

Millman: I keep being put in an awkward position, because I basically want my example to be such that the pretest scores for the two groups won't differ from the corresponding post-test scores, and that is the problem. Julian has returned to a situation where that is not the case, and we are on different wavelengths again.

Stanley: I hate to do this thing in our heads, but let's try it. Suppose we have the ten IQ scores I talked about. Okay, we have them and we know the reliability. For each obtained score, we then get an estimated true score, with a reliability of .8 and a regression 20% of the weight of the mean in both directions. We then look back at the obtain scores again, and assign five children to one group and five to the other. Then, we will look at the estimated true scores, because they are our best estimate of what will happen on the retest. We don't really need the retest directly. We have to have the retest indirectly, in order to get the reliability coefficient. Okay, we are pretty sure that the mean of the estimated true scores will be higher for the high group than for the low group, even though the low group has moved up some, and the high group has moved down some, but they haven't necessarily moved all the way up or down.

Millman: Right. Let us pretend that in your group of ten, there are two people who have the same observed score. It turned out that the one got in with the high group and the other in the low group. When we come along to matching, these two people get matched. There is no reason to expect regression.

Stanley: You are perfectly right. If two people have exactly the same obtained score, on the same test, at the same time, then the estimated true scores will have to be the same for those two people. They cannot be otherwise.

Millman: You have been having trouble identifying with me because you have not treated the two groups as coming from the same original population.

Stanley: Well, regression always applies to an ex post facto observation of a set of obtained scores. There always has to be ex post facto analysis. You have got to obtain the scores, and look at them after they have been obtained, to see who is high relative to the group and who is low. One can't talk about in it any other fashion. Suppose you put inherently different subgroups together as one big group. The reliability coefficient is increasing a lot because error measurement is really a property of the individual, whereas the variability is a property of the way the population is put together, and the bigger the variability, in effect, the bigger the reliability. Now, the amount of regression will go down if one chooses to put Blacks and Whites and so forth, together in a huge, extended population. The estimates of the true scores will be less discrepant from the obtained scores, than they would be within any one of the subpopulations.

Hopkins: I don't see that. I am looking at the difference in the obtained score and the predicted true score. I don't see how regression is going to be changed by group heterogeneity.

Stanley: Well, look at it this way. You know that the reliability coefficient is a function of 1 minus the variance of errors of measurement over the variance of obtained scores. Now, the variance errors of measurement is not affected by how we happen to group people very much, unless we throw in different kinds of highly variable people. We can really blow up the appearance of obtained IQ scores radically so that whereas for Blacks alone we might have a reliability coefficient of .7, for Blacks and Whites put together we might have a reliability coefficient of .9. The regression equation we use is a very simple one, in that it is only the mean of the whole population, plus the reliability coefficient, times the obtained score, minus the mean of the whole population. It does not include the variability component at all. The regression is bound to be less, because the regression is entirely, in terms of that reliability coefficient.

Hopkins: We have increased the true variance, too.

Stanley: That is what makes the reliability go up—sure. The true variance is increased a great deal by having this quite heterogeneous population ranging from perhaps 60 to 130, or so. But the thing is, the estimated true score formally has no variance in it.

Hopkins: For obtained scores, the regression parameter seems to me to be unrelated to group heterogeneity.

Stanley: This is not true for the estimated true score. We just have two different things: reliability coefficient versus the variance error of measurement. They do two different jobs for us, as we know. I don't seem too confident about it, because you and I aren't asking the same question. It is clear that the reliability coefficient has to go up. Okay now, if it is true, then I can derive what I have been saying, easily, because I did it in my reliability chapter, half a dozen times. If you make the group heterogeneous enough then the obtained score and the estimated true score will be the same. Virtually the same.

Mann: For a while I though we really had it, and something slipped out of our hands, again.

Stanley: I think we tend to deal with two kinds of regression when we use the word. In one case, we think of the x - y regression, where x and y are both obtained scores, and maybe even obtained scores in different functions. In another case, we are talking about regression due to errors of measurement, alone.

Millman: Let me start over. The context for this was a special education researcher who has two groups of people that he is dealing with, and these groups have not been randomly assigned. The two groups are assumed to have come from a common population. The investigator has pre-treatment information about them, and he wants to now assess the effectiveness of some kind of intervention. Then, he has some sort of post test results, or information, and the question is how might he analyze those data? Two techniques were suggested: First, match the students from the two groups on that pretest information; second, use analysis of covariance in which the pretest measures are used as covariates for adjustment.
Both of those techniques can result in misleading information, because of the regression effect. That is, some differences between the two groups might be expected, even if there were no real treatment effects at all. However, the problem that we have had in this long discussion concerns what are those situations where regression effects do not always occur. That is, there are situations in which you don't have expected regression effects. Then the problem becomes, what are those situations in which you can expect regression effects, and what are those situations where you can't. You could put a Figure 1 and Figure 2 in this thing. When the situation is depicted as in Figure 1, there is no regression effect. In Figure 1, there is. It's not a simple thing, but, in general, the more that the two groups, that you are comparing, are distinct populations in their own right, there is a good probability that you are going to have a regression effect.
If you have a group of Down's Syndrome people and a group of normals that are your tow groups, then almost assuredly you

would expect these to be considered distinct populations, and you must be aware of the regression effect. It is a pretty tricky thing, I think—it is easier to get agreement, case-by-case, on when the regression effect is likely to occur, than it is to make a definitive statement about when regression will not occur. To elaborate a little, if you had Blacks and Whites, I don't think there would be any question that the regression effect would be very possible. If you had students in parochial schools and students in private schools, it might be. If, however, you have students in Miss Jones' class across the hall and Miss Smith's class right here, and the two groups haven't been all that differently composed, then regression effect might not occur. That might not be as serious.

Hopkins: I think Jay has done a very good job of stating it.

Stanley: May I make a caution here, now. In terms of covariance analysis, we have to keep in mind, eternally, that in the so-called true experiment, you don't have this regression type problem. We have a problem of weak covariance, maybe. But, we don't have the problem of bias in a true experiment. On the other hand, in the true experiment, we are better advised, in certain cases, to use randomized blocks than to use covariance, but we have to consider too many degrees of freedom we will lose with randomized blocks. If there are very few treatments—say, two treatments—we are wiping out half the degrees of freedom by using blocking, so we lose a lot of power, and in a small experiment we can't afford to lose our power. Here, we have to try to use covariance analysis in a ture experiment. Cox, in his 1957 article, shows that randomized blocking is superior to covariance for most purposes, but only in the type one error sense, not in the parlor sense.

Millman: Isn't there also something to be said then, in terms of beta?

Stanley: Yes, in terms of alpha and beta, it is a trade-off. But, now suppose one has sixteen subjects—eight in each group— and you randomly block. Instead of 14 degrees of freedom, you now have seven for the randomized block design. Well, 13 for covariance and 7 for randomized blocking. You've lost almost half the degrees of freedom. Now, if his reduction in error variance by blocking is that much greater than reduction by covariance, he'll win. But, it probably won't be. He's probably lost appreciable power and, therefore, can only detect a much bigger difference than he used to detect. On the other hand, if he has 100 cases, and he reduces them to 49 degrees of freedom, well, he's not in such bad shape, probably. He's probably well advised to use blocking. One disadvantage of blocking, that is often forgotten, is that it requires you to keep the experiment on the rails, but covarying in effect doesn't. To covary, you just get your antecedent measures, do your experiment, and analyze. In blocking, you set up a design in which you run into the problem of lost observations, which can be pretty serious. You'll have to have a whole experimental strategy, to avoid lost observations if you block, but if you covary, it is much less of a problem.

## REFERENCES

Evans, S. H. Reply to Sprott. Psychological Bulletin, 1973, 79, 180.

Glass, G. V., Peckham, P. D., & Sanders, J. R. Consequences of failure to meet assumptions underlying the analysis of variance and covariance. Review of Educational Research, 1972, 42, 237-288.

Sprott, D. A. Comment on misuse of covariance. Psychological Bulletin, 1973, 79, 180.

## CHAPTER 21

# RESEARCH DESIGN AND METHODOLOGY:

# RESULTS OF AN OLDER CONFERENCE

Overview: In 1959, a number of leading researchers and practitioners in the area of mental retardation met with non-retardation-oriented design and statistical specialists to clarify issues in retardation research. The participants met in Philadelphia at a 3-day meeting sponsored jointly by the Woods School in Langhorne, Pennsylvania and AAMD. The invited guests prepared formal papers beforehand for the Woods Conference. In contrast to the much less rigid structure of the 1972 Montgomery Conference. The entire September, 1959, issue of the American Journal of Mental Deficiency devoted its 207 pages to highlights of the Woods Conference. It is of historical interest to see how the pholosophy of, and technology of, research with the handicapped have changed over thirteen years. For this reason, the present chapter excerpts selected ideas from various papers given at the Woods Conference.

The preceding chapters have dealt with the more or less spontaneous "happenings" in a research design and methodology conference held jointly by the Montgomery Intermediate Unit No. 23 and the Pennsylvania Bureau of Special and Compensatory Education in February, 1972. The editors of this volume feel the format of the three-day meeting lent unique advantages to the content. First, none of the participants were asked to contribute formal papers. Instead, the participants merely had to show up with a clear head and be prepared to do some fast mental footwork on some of the issues involved. Second, while there was a tentative list of topics to be covered at the Montgomery Conference, the flexibility was always present whereby one could diverge into whatever areas he wished. Third, because the participants did not have to concern themselves with presenting and defending a formal paper, they could render intuitive, down-to-earth insights that might otherwise not find their way into print. Fourth, because the conference was informal in nature, the participants felt at home in criticizing each other whenever appropriate and at commenting, however often they wished, on a single issue. In summary, the difficulties of attempting to organize the transcripts of the Montgomery Conference into topical chapter threads appear to have been more than offset by the quality of highly readable, intuitively understandable insights provided. While the Montgomery Conference appears to be a first of its kind, there does appear to be one other research meeting whose findings should be described: the Woods Conference.

On May 1, 2, and 3, 1959, a "Symposium on Research Design and Methodology in Mental Retardation" was held at the Woods School, Langhorne, Pennsylvania. An issue of the American Journal of Mental Deficiency (Volume 64, Number 2), was devoted to the proceedings from this conference. The distinguished participants were: James J. Gallagher, Lawrence M. Stolurow, Lyle V. Jones, Sidney Siegel, Leon Festinger, Irving Lorge, Margaret Mead, Jerome Cornfield, Boyd R. McCandless, Herbert G. Birch, Richard L. Masland, Benjamin Pasamanick, Harvey F. Dingman, Nicholas Hobbs, Rue L. Cromwell, Morton Kramer, Herbert Goldstein, Dee W. Norton, Thomas Gladwin, Dorly D. Wang, and Theodore Greiner. This research conference was one of the first large-scale attacks on the issue of what is "special" in special education research. Too often, it is assumed that one who has questions about statistics and research design can simply consult the existing texts on these subjects. Unfortunately, almost all such texts were written by the general educational researcher for general education research, where it is assumed that one does not have small sample problems and so forth. There are a host of content-specific design and statistical problems inherent in almost any kind of special education research. Until recently, most educators—general as well as special—blithely assumed such problems did not exist. Conferences such as that held at Woods School thus serve a valuable function in dispelling these delusions.

In attempts to establish historical footings for the Montgomery Conference's efforts in recognizing problems in experimental design and statistics that are unique to the realms of special education, one must ask what the Woods Conference found. Because not all of the conference was directed at statistical and experimental design problems, there is a large amount of discussion, on the one hand, of previous directions of research with regard to program content and resultant findings, and, on the other hand, of where research should be directed in the future. Thus, to a great degree, the conference was directed at the whole issue of mission-oriented research. Further, the content area of special education consisted only of retardation and not the other, equally important areas such as emotional disturbance, speech impairment, learning disabilities, etc. Nonetheless, there are many generalizable thoughts resultant from the conference on statistics and experimental design.

Many papers in the Woods Conference were directed at very specific, non-generalizable topics (e.g., monitoring patient flow in a hospital, etiology, nomenclature). Other papers, however, yield some generalizable insights into the nature of research with the mentally retarded. The studies can be classified into about six major categories: (*a*) design and statistics, (*b*) testing, (*c*) miscellaneous design considerations, (*d*) data bank activities, (*e*) social aspects of research, and (*f*) program evaluation. Selected ideas from each study are presented below.

*Design and Statistics*: Norton (1959) talks about the fact that while design and statistical analysis must be appropriate and indeed can become sophisticated, such efforts are wasted if adequate conceptualization of the problem and the theory behind it has been slighted. Norton is also concerned about the nature of sampling in mental retardation research. The fact that most retardation researchers do not sample randomly from large populations demands that research writers describe much more specifically the nature of their subjects. Another aspect of sampling that bothers Norton is the lack of replication of a diagnostic blocking variable in a study. He feels that diagnostic group differences should be measured under several conditions within a study to maximize generalizability. With regard to randomization in sampling, Norton feels an extraneous variable can be controlled effectively to the extent that sample size is adequate, and to the extent that the variable is correlated with the criterion. Finally, Norton gives some sound advice in pre-experimental planning that unfortunately is rarely followed: "First, decide on non-statistical grounds just how large the true difference between population means would have to be in order to justify either a practical decision or an intensive theoretical follow-up. Second, specify what odds you would like to have, of being able to detect this difference in a particular experiment. Third, decide on the risk you are willing to take of rejecting a true hypothesis, i.e., establish your definition of a rare event. Fourth, develop an estimate of the variance of the treatment populations. A value—at least bracketing values—can often be secured from previous researches, or from analysis of the possible range of the criterion scale, or, if necessary, from a pilot study. Fifth, and last, working backwards from the formula for the test statistic to be used, estimate the approximate sample size required. Given an initial estimate of the size required in a simple-randomized experiment, one is then in a position to consider the desirability of increasing sample size, or of improving the reliability of the criterion measure, or of switching to a design in which precision is increased by taking account of variables related to the criterion measure (pp. 376-377).

Siegel (1959) discusses four criteria that the experimenter can use in selecting schemes of statistical analysis. The first criterion is the nature of the population distributions of the variables under study. Siegel suggests that if the variables do not appear to be normally distributed, them some nonparametric technique might be appropriate. The second criterion concerns the nature of the data. Depending upon whether the data are in the form of frequencies, ranks, or numerical scores, certain types of analyses are more appropriate than others. With regard to the power of the data, Siegel prefers numerical scores to ranks, which in turn are to be preferred to frequencies. Third, Siegel suggests that the nature of the questions asked in the research determines, in large part, what specific variations in a given statistical technique are to be used. In this regard, Siegel claims that if directional hypotheses cannot be stated about the research problems of interest, then adequate conceptualization of the study has not occurred. Finally, Siegel suggests that the experimenter consider the power of the statistical test: "A statistical test is considered a good one if it has small probability of rejecting the null hypothesis when it is true, but a large probability of rejecting the null hypothesis when it is false and thus should be rejected (p. 402)." Parenthetically, it is of interest to note that Siegel considers the present status of mental retardation theorizing to be inadequate for postulating higher order interactions.

Testing: A number of papers at the Woods Conference dealt with measurement. Jones (1959) provided a general overview on various testing problems. He claimed that there is "a risk that students of test theory attend to the mechanics of test construction to the exclusion of the imaginative, creative processes required to develop adequate measures for difficult research problems (p. 390)." Further, Jones has serious reservations about the use of developmental measures in mental retardation research, such as standardized IQ tests. He suggests that the factor analytic structure of sets of items changes markedly as one gives them to progressively older children. One does not know the relative weights of the various factors that make an IQ score at any given age; e.g., is the memory span factor more or less important to total IQ score at age 10, compared to age 13. Because of the unknown weighting of the many factors that comprise total IQ, Jones claims a standardized IQ test is not one test, but "rather a series of tests, not identical or even parallel in content (p. 385)." He suggest turning to other types of measures.

Gladwin (1959) posits that most currently available standardized intelligence and achievement tests do not contain the types of behavior that are useful in diagnosing specific learning problems; the test content deals only with more global payoff tasks. On the other hand, Gladwin feels teacher expertise and intuition do far better than standardized tests in determining where next to give instruction to a child on the curricular continuum. Gladwin (1959, p. 321) further asserts that almost no information has been gathered on what is and what is not a prerequisite for successful functioning at any given task in the curricular hierarchy.

Stolurow (1959) extends Gladwin's concerns about practical educational programing for mentally handicapped children on the basis of curricular prerequisites. Stolurow suggests that while researchers have looked intensively at general intellectual deficits of the retarded, educational content-area deficits have not been touched upon. Further, Stolurow suggests that while basic mental retardation research is usually based upon sophisticated theory, the more typical applied (or fundamental) research should rely upon a careful task analysis of learning hierarchy prerequisites.

Turning away from intellectual and educational testing, Gallagher (1959) considers the problems of personality measurement in younger retardates. While intellectual and social measurement has progressed nicely in mental retardation research, the development of personality measures has been wanting. Gallagher discusses weaknesses in current personality assessment schemes. First, objective self-report procedures are questionable because of reading difficulties and self- misperceptions. Second, projective devices often lead to invalid protocols because of inadequate verbal production, temporal fluctuations, and stimulus boundedness. Third, observation methods are peculiarly subject to unrepresentative, temporal fluctuations of the child. Thus, Gallagher and his associates went ahead and developed a rating scale based upon the expert informer type of respondent. Details of the scale are given.

Cromwell (1959) also had several thoughts on personality measurement with the retarded. He suggests that while retardates differ in obvious ways from normals on intellectual variables, it is possible that on the other hand, personality mechanisms may be quite similar. Cromwell describes a theory of generalized expectancy to help explain personality features. "The component of expectancy...develops as a function of the individual's successes and failures in past situations in his life and... generalizes into the present situation. The higher the ratio of successes to failures, the higher the generalized expectancy (p. 333)." Cromwell goes on to describe a vertical lift apparatus found useful in studying the implications of this theory.

Miscellaneous Comments About Experimental Design: Apart from the studies mentioned above from the Woods Conference that reflected, more or less, common themes, a number of papers contained isolated design comments that bear repeating here. Cornfield (1959) began his paper with a catchy little anecdote: "On being asked to talk on the principles of research, my first thought was to arise after the chairman's introduction, to say, 'Be careful,' and to sit down. On hearing my more extended remarks, you may regret my not having followed my first impulse (p. 240)." Of interest are Cornfield's suggestions to use restricted (stratified) randomization to increase the sensitivity of an experiment and to perform replications of studies wherever possible (a sorely lacking feature of most mental retardation research).

In response to the question as to what is different in research methodology for the handicapped, relative to normal children, McCandless (1959) takes a strong position that there are no essential differences.

However, the results of the Montgomery Conference would suggest that while this is true by the letter, it certainly is not true in spirit. McCandless also questions the appropriateness of matching techniques in mental retardation research whereby one matches retardates with CA "X" and MA "Y" with either normal children of CA "X" or CA "Y."

Siegel (1959, p. 396) is quite concerned with the dearth of true theory in mental retardation research. He does not see how any research can proceed without theory.

Festinger (1959) is also disturbed about the issue of theory in retardation research. In fact, he was puzzled as to why a conference on methodology would be called when theory itself in the area of mental retardation was not yet established. Festinger discusses very lucidly the overemphasis of researchers upon sophistication in design and statistical analysis while they almost completely ignore sound theorizing. In terms of sampling, Festinger is disturbed by analysts who derive relationships on the whole sample, which usually contains pre-identified, widely disparate subgroups. He would rather see investigators test relationships within each meaningful subgroup, and then to look for consistency across subgroups. While Festinger would have the experimenter concentrate on sound conceptualization and theorizing, he does grant (1959, p. 369) that poor reliability of criterion instruments could lead to a finding of little, or no, difference between treatment groups, whereas in reality, there might be a quite large difference.

Several speakers at the Woods Conference brought up the topic of observation. Of particular note is Birch, who claimed: "There is no observation independent of theory. The question is whether your theory is right or whether your theory is wrong, whether you are observing what occurs or whether you are projecting what you wish to observe (1969, p. 302). Further, Mead (1959, p. 303) suggests most observational systems are inappropriate samples of the total behavior occurring. Instead, she recommends sound moving pictures to capture all details in a given situation.

Data Bank: As more researchers became aware of the need for replicated, well-conceptualized research, the need for data banks such as Project TALENT arose. However, data banks for handicapped children have been nonexistant. To meet this need, Lorge (1959) issues a plea for research agencies, institutions, and schools to pool common types of data on the children with whom they deal. Lorge describes, in exhaustive detail, a rather comprehensive coding system for research data derived from retarded children. The major categories in his system are: etiology, intelligence (total MA, total IQ, verbal MA, performance MA, educational achievement, specific abilities), maturation (self-help, motor development, socialization), psychosocial status (emotional adjustment, social adjustment, temperament), physical and environmental status (sensory handicaps, motor handicaps, speech handicaps, environmental status).

Social and Ethical Aspects of Research: Because of the dangers associated with many drugs and the questionable propriety of withholding treatment from one or more groups of children, social and ethical issues of research have come into prominence. Greiner (1959) offers several observations in this area. First, he suggests: "In years to come, the retarded may claim an all-time record, of having the greatest variety and largest tonnage of chemical agents shoveled into them (p. 346)." He goes on to point out that a number of clinical drug studies from medical research are so flawed that it is dangerous to draw conclusions from them. Because of these dangers, double-blind drug studies are called into serious question.

Hobbs (1959) discusses a number of fascinating behind-the-scenes issues about research. In defense of efforts such as the Woods Conference and the Montgomery Conference, he says: "I think it might be very fruitful to program a series of small research conferences, to bring together scientists whose interests intersect at the point of junction between some aspects of the problem of mental retardation and some related area of inquiry. whose explorers would doubtlessly be shocked to get an invitation to such a conference, but who might well have a germinal influence on the field of our concern (pp. 231-232)." Hobbs also takes note of the necessity for administrators to arrange for conditions within their programs that are conducive to conduct of research. Researchers should not, for example, be straddled with the dual responsibilities of research and service functions. Also, researchers need a lot of time to spin their mental wheels; overt evidence of continual production by a researcher will not always be in view.

Program Evaluation: At the time of the Woods Conference, accountability had not yet gotten up a head of steam. Thus, it is of interest, historically, to hear what early thinkers in this area had to say. Goldstein (1959) emphasized the need to evaluate, in formal ways, the effects of training programs for retardates. He was particularly concerned about the poor descriptions of techniques in research and evaluation reports that prohibited replication of a study.

## REFERENCES

(Note: Because all the papers listed below were contained in Volume 64, 1959 of the American Journal of Mental Deficiency, only the authors, titles, and pages are given, to avoid redundancy of detail.)

Birch, H. G. (Discussion of papers), p. 302.

Cornfield, J. Principles of research, pp. 240-252.

Cromwell, R. L., A methodological approach to personality research in mental retardation, pp. 333-340.

Festinger, L. Sampling and related problems in research methodology pp. 358-366.

Festinger, L. (Discussion of papers), p. 367.

Gallagher, J. J. Measurement of personality development in preadolescent mentally retarded children, pp. 296-301.

Gladwin, T. Methodologies applicable to the study of learning deficits, pp. 311-315.

Gladwin, T. (Discussion of papers), p. 321.

Goldstein, H. Methodological problems in research in the education programs for the treatment of the mentally retarded, pp. 341-345.

Griener, T. Problems of methodology in research with drugs, pp. 346-352.

Hobbs, N. Research in mental retardation: Prospects and strategies, pp. 229-239.

Jones, L. V. Problems of devising and selecting appropriate measurement tools, pp. 384-394.

Lorge, I. Planning cooperative utilization of data on the mentally retarded, pp. 269-277.

McCandless, B. R. Measurement problems in research, pp. 265-268.

Mead, M. (Discussion of papers), p. 303.

Norton, D. W. Problems in experimental design, pp. 370-383.

Siegel, S. Problems in analysis of data, pp. 397-405.

Siegel, S. (Discussion of papers), p. 396.

Stolurow, L. M. Requirements for research on learning in mental deficiency, pp. 323-332.

## TOPICAL INDEX

### EXPLANATION

For maximum usage of the Topical Index, we offer a clarification and examples of the format.

The discussants' names have been abbreviated:

| | | | |
|---|---|---|---|
| Bracht | (BR) | Hopkins | (HO) |
| Cross | (CR) | Mann | (MA) |
| Feldhusen | (FELD) | Millman | (MILL) |
| French | (FR) | Proger | (PRO) |
| Hagin | (HA) | Stanley | (STAN) |
| Entire page | (ALL) | | |

A hyphen between names denotes sequence; a comma, unrelated discussion in between the referenced topic; the word "to" indicates two or more sequential pages of related discussion and implies inclusion.

Fast, accurate identification of a referenced topic is provided by the major sectioning of each page into top, middle (mid), and bottom (bot), with extended minor section of "above mid" and "below mid" occurring on pages of multiple, varied discussions.

The following examples take into account the total format:

| | | |
|---|---|---|
| Blocking | 32 mid (FELD) to 33 (FELD); | Page 32 is comprised of nine separate dialogues—with (FELD) appearing at the middle and bottom—the referenced topic of "blocking" begins with the middle entry. The discussion is carried through sequentially to page 33, where (FELD) is entered only once, and concludes with this entry. |
| | 96 (FELD) to 97 (ALL); | The referenced topic of "blocking" is again picked up with the single entry (FELD) on page 96 and is continued without interruption throughout all of page 97. |
| | 208 (FELD) | Topical discussion of "blocking" occurs with this single (FELD) entry on page 208. |
| | 283 (STAN)-(HO); | This entry illustrates the use of the hyphen—sequential, contiguous discussion of the |

| | |
|---|---|
| | "blocking" topic—first by Stanley and immediately after by Hopkins. |
| 296 mid, bot (STAN) | This double (STAN) entry—separated, in this instance, by a single line of dialogue (here, Millman was speaking on a different topic) illustrates the recurring separations of the more-detailed, focal discussions from the "leading up to" and superfluous dialogues, as well as, in some circumstances, interjectory unrelated material. |
| Cell Frequencies in a Factorial Design<br>214 above mid (HO) to 215 (ALL) | On page 214 there are four (HO) entries, requiring more than usual sectioning of the page to pinpoint the topic. |

Ability Training 57 mid (MA); 68 (STAN); 69 (STAN); 76 mid (MA)-(STAN); 77 (MA)-(STAN)-(FELD); 80 mid (BR), bot (MA); 81 (STNA); 82 bot (FELD); 160 mid (HO).
Ability vs. Talent 79 top (STAN).
Accountability 106 mid (MA)-(MILL)-(HO)-(PRO); 112 (FELD); 118 (MILL); 119 (MILL)-(MA)-(MILL); 120 bot (MILL); 121 bot (PRO).
Achievement 166 (CR).
Addition 160 bot (BR).
Adjusted Data 260 bot (STAN).
Administrative Problems 213 top (MILL), (FR).
Affective Instruction 164 top (MILL).
Affective Tests 113 (MILL).
Analysis of Covariance 209 (HO) to 210 bot (FR); 216 top (MILL)-(FELD)-(MILL)-(HO); 218 above mid (MA)-(HO)-(MA)-(HO)-(MA); 260 bot (MILL), bot (STAN); 261 top (HO). See also Covariance.
Analysis of Variance 194 (STAN); 260 mid (STAN), below mid (STAN); 261 top (STAN); 270 top (MA); 271 top (STAN); 272 (MILL).
Anonymity 184 top (FR)-(MA)-(FR)-(HA)-(FR).
APA Standards 131 top (HA); 132 mid (PRO)-(MA).
Applied Research and Technology — see also:
  Ability Training
  Accountability
  Administrative Problems
  Affective Instruction
  Aptitude Treatment Interaction
  Arithmetic
  Behavior Modification
  Certification
  Chemistry
  Computer-Assisted Instruction
  Conditioning
  Cost Data
  Cost Effectiveness
  Curriculum Evaluation
  Curriculum Package Results
  Creativity
  Data Banks
  Diagnostic Prescriptive Teaching
  Discovery Learning
  Dissemination
  Effective Treatments w/non-ATI Implications
  Engineered Classroom
  Faculty Psychology
  Formative Evaluation
  Grading
  Guided Discovery
  Individualization
  Infant Education
  Instruction Evaluation
  Instructional Materials
  Instructional Materials Research and Development

  IQ Training
  Mastery Learning
  Materials Research and Development
  Medical Results
  Microteaching
  NAEP
  Normalize
  Operant Conditioning
  Performance Contracting
  Permanent Psychological Records
  Precision Teaching
  Prediction of Vocational Success
  Preschool Programs
  Prescription Teaching
  Prescription Validation
  Program Evaluation
  Project TALENT
  Rap Psychology
  Remediation
  Summative Evaluation
  Science
  Self-Help Skills
  Social Studies
  Task Analysis
  Teacher Effectiveness
  Team Teaching
  Training Objectives

Aptitude Tests 71 (STAN); 159 bot (STAN).

Aptitude Treatment Interaction 15 (MA), (FELD); 30 (ALL) to 31 below mid (BR); 32 mid (FELD) to 34 (MA); 35 top (BR) to 37 (ALL); 38 (ALL); 57 mid (MA); 64 (MILL), bot (MA), bot (FELD); 67 (STAN); 80 top (HO), (BR), (HO).

Arithmetic 71 (STAN); 159 bot (STAN); 160 bot (BR); 162 top (BR); 163 (MILL)-(HO); 164 (STAN); 165 (BR); 169 top (FR).

Arithmetic Disability 24 (BR).

Art 159 bot (STAN).

Attenuation of Correlation 255 bot (MA) to 257 (ALL).

Attrition 271 bot (STAN).

  Attrition - Biased 270 bot (FELD); 271 top (STAN).

Attrition - Natural 271 top (STAN).

Bayesian Statistics 63 (FELD); 185 (MILL); 199 top (MILL); 200 top (MA)-(MILL)-(MA)-(HO); 274 (MA)-(STAN), bot (MILL); 275 (STAN).

Behavioral Modification 58 (FELD)-(MILL); 59 (MILL); 62 (MILL); 65 (MILL). See also Conditioning, Operant Conditioning.

Behavioral Objectives — See Objectives, Objective Hierarchy.

Bias 273 (ALL). See also Contamination

Bias Contamination 284 (STAN).

Blind 141 top (HO); 229 (HO)-(MILL).

Blocking 32 mid (FELD) to 33 (FELD); 96 (FELD to 97 (ALL); 208 (FELD);

214 top (MA) to 216 (ALL); 229 (HO); 270 top (PRO); 283 (STAN)-(HO); 296 mid, bot (STAN).
Bloom's Taxonomy 81 (STAN). See also Objectives, Objective Hierarchy.
Brain Injury 95 bot (MA) to 96 (MA).

Case Study 58 (FELD); 59 top (FELD), (STAN); 60 below mid (STAN).
Case-Study Approach 31 below mid (BR); 40 (HA); 197 top (FELD)-(HO); 198 bot (HO).
"Cash Validity" 12 (FR). See also Validity.
Cause-Effect Relationship 84 (HOO. See also Control.
Ceiling Effects 261 bot (STAN); 262 mid (STAN); 277 bot (MILL).
Ceiling Tests 277 (STAN).
Cell Frequencies in a Factorial Design 214 above mid (HO) to 215 (ALL).
Cerebral Palsy 9 (FR), 41 mid (FR), 132 top (FR).
Certification 125 bot (MILL).
Change 260 above mid (MA); 164 (ALL) to 267 (ALL). See also Gain.
Chemistry 168 bot (STAN).
Cognitive Tests 113 (MILL).
Commercial Systems 169 (CR)-(FR).
Comparative Studies 54 (FELD).
Computer 287 (HO).
Computer Assisted Instruction 65 bot (HO); 66 (FELD)-(MILL); 67 top (FR).
Conditioning 237 mid (STAN); 262 mid (STAN)-(MILL); 266 above mid (STAN). See also Behavior Modification, Operant Conditioning
Confidence Intervals
Constructs 166 (CR). See also Validity, Reification of Constructs
Construct Validation 130 top (PRO); 134 (PRO), (BR); 137 (BR), bot (MILL). See also Validity.
Contamination 60 bot (FELD); 62 (STAN); 237 mid, bot (STAN).
Contamination and Bias 197 top (FELD) to 199 (BR); 200 bot (FELD) to 201 top (FELD); 218 top (MA) to 220 (ALL). See also Bias.
Content Validation 131 bot (HA); 133 (MILL); 134 (PRO)-(MILL), (BR); 135 (MILL); 136 (BR); 137 (BR), bot (MILL); 160 mid (HO); 162 (MILL). See also Validity.
Contradictory Results 180 (FELD); 181 mid (FR).
Control Group 88 (BR) to (MA); 89 (HO); 110 (HO); 111 top (MILL); 196 (BR); 229 (HO)-(MILL); 230 bot (MA) to 235 (ALL).
Controlled "One-day" Studies.
Correlation 54 (HO); 84 (HO); 179 (FELD).
Cost Data 113 (MILL).
Cost Effectiveness 107 top (HA); 108 top (MA) to 109 (MA).
Covariance 236 top (STAN); 237 bot (STAN) 283 top (MA), (STAN); 284 (STAN); 286 mid and bot (MILL); 289 top (HO); 296 mid and bot (STAN).
Creativity 165 (MILL). See also Reliability.
Criterion-Referenced Measurement 20 bot (MA) to 24 (FELD); 54 (FELD); 64 (MILL); 70 top (BR); 159 top (MA) to 171 (ALL); 196 (BR). See also Domain-Referenced Measurement.
Critical Reviews 132 (FELD)-(MA), below mid, bot (FR); 136 (FELD).
Cross Validation 277 (STAN). See also Validity.
Cultural Bias Tests 170 (MA). See also Bias, Contamination.

Cultural Deprivation 9 (FELD); 285 (BR).
Curriculum — See also:
    Ability vs. Talent
    Addition
    Affective Instruction
    Arithmetic
    Art
    Behavior Modification
    Bloom's Taxonomy
    Chemistry
    Conditioning
    Commercial Systems
    Computer Assisted Instruction
    Creativity
    Discovery Learning
    Discrimination Learning
    Engineered Classroom
    Faculty Psychology
    Guided Discovery
    Infant Education
    IQ Training
    Objective Hierarchy
    Objectives
    Operant Conditioning
    Performance Contracting
    Preschool Programs
    Rat Psychology
    Reading
    Reading Instruction
    Science
    Self-Help Skills
    Sequence of Instruction
    Social Studies
    Task Analysis
    Team Teaching
    Theories of Learning
    Training
    Training Abilities
    Training Intelligence
    Transfer
Curriculum Evaluation 112 mid (MA); 113 (ALL). See also Program Evaluation.
Curriculum Package Results 121 (PRO). See also Instructional Materials.

Dangers 229 (MILL).
Data Banks 64 top (MA)-(MILL); 65 (MILL)-(HA), (HO); 67 mid (FR); 68 (MA); 69 bot (STAN); 71 (HA).
Data Coding Errors 75 (FELD).
Data Deletion 271 bot (STAN).
Data Estimation 273 (HO); 274 (HO).
Data Pooling 185 above mid (FR)-(MA); 286 top (FELD)-(MILL)-(FELD)-(HO). See also Small Sample Size and Replication.

Data Reduction 81 (FELD)-(STAN); 277 (FELD); 278 (HO).
Defining Problem 101 (ALL) to 103 top (MILL).
Degrees of Freedom 296 mid, bot (STAN).
Descriptive Research 20 bot (MA) to 24 (BR); 31 (PRO) to 33 (FELD); 34 (HO); 201 mid (FELD).
Descriptive Statistics 83 bot (STAN); 194 top (MA)-(BR); 195 (BR), (FELD); 196 (FELD), (BR); 278 (HO).
Developmental Studies 259 bot (MA).
Diagnosis 9 (FR); 96 (FELD) to 97 (ALL).
Diagnostic Interviews 164 (STAN); 166 (HA).
Diagnostic/Prescriptive Teaching 58 (FR); 168 (CR).
Diagnostic Tests 71 (STAN); 159 top (FR)-(ALL); 161 mid (BR); 164 (STAN).
Difference Indices 160 mid (HO).
Discovery Learning 71 (ALL).
Discrimination Learning 162 (MILL).
Dissemination 187 (PRO), (FELD); 188 (MA) to 189 (ALL).
Disturbed 62 (STAN). See also Emotionally Disturbed.
Domain-Referenced Measurement 162 (MILL).
Down's Syndrome 259 above mid (MA) to 260 mid (MA), below mid (STAN); 261 bot (STAN); 266 mid (STAN).

Educable Mentally Retarded 55 top (PRO); 57 top (MA); 68 (STAN); 215 bot (PRO)-(MILL); 165 (MA).
Effective Treatments w/non-ATI Implications 40 (MILL) to 44 (BR); 47 (MILL) to 49 mid (MA).
Emotionally Disturbed 9 (FELD); 65 mid (HA); 96 top (HA); 210 bot (PRO); 212 mid (FELD); 218 below mid (MA)-(ALL). See also Disturbed.
Engineered Classroom 62 top (HA).
Errors of Inference 211 (HO).
Estimated True Score 283 (STAN); 284 (STAN).
Ethics of Experimentation 230 bot (MA) to 231 (FR); 232 mid (FELD); 233 (HO).
Ethnic Groups 145 (HA)-(HO); 146 (MA); 147 (CR); 164 (STAN); 170 (MA),(MILL), bot (STAN); 259 mid (STAN); 262 bot (STAN); 263 top (MA); 286 top (MILL); 288 top (HO) to bot (MILL); 294 mid (STAN); 295 (MILL).
Expectancy Effects 88 (BR) to 89 (HA), (FELD); 91 (FELD) to 92 (ALL); 94 bot (HO) to 97 (ALL).
Expectations 174 bot (MILL); 275 (STAN).
Experimental Editions 131 top (HA).
Experimenter Bias 181 below mid (FR).

Factor Analysis (SOI) 76 top (MA); 80 mid (HO) to 82 (ALL); 83 bot (STAN); 201 mid (FELD); 176 (MA) to 279 (ALL).
Faculty Psychology 76 mid (MA).
Floor Disadvantage - Test 277 (STAN).
Floor Effects 261 bot (STAN); 262 mid (STAN); 276 (STAN); 277 (MILL).
Formative Evaluation 58 (FR); 113 (STAN); 194 (BR); 195 (BR).
Fractionation of Assessment 76 (FR); 77 (FELD); 78 (HA); 79 top (STAN); 81 (STAN).

Gain 236 top, bot (STAN); 237 mid (STAN); 259 bot (STAN); 260 top (MA), above mid (STAN)-(MA)-(MILL), below mid (STAN); 261 (HO), bot (STAN);

284 (STAN); 287 (MILL). See also Change.
Generalizability 35 bot (BR) to 36 (FELD); 44 (PRO) to 46 (BR); 185 mid (FR); 194 (STAN); 195 (BR), bot (HO); 196 (MILL), 211 (BR), 215 (MILL); 216 bot (HO).
General Linear Hypothesis 270 top (STAN).
Genetic Conditions 87 top (HO).
Giften 62 (STAN) to 64 (STAN); 67 (STAN); 48 (STAN); 71 (STAN): 79 bot (STAN); 95 (FELD)-(MA); 164 (STAN); 190 bot (STAN).
Grading 160 mid (HO).
Guided Discovery 70 top (BR).

Handicapping Conditions and Special Groups — See also:
  Blind
  Brain Injury
  Cerebral Palsy
  Creativity
  Cultural Deprivation
  Disturbed
  Down's Syndrome
  Educable Mentally Retarded
  Emotionally Disturbed
  Ethnic Groups
  Genetic Conditions
  Gifted
  Heredity
  Labels (See "Sampling")
  Labeling
  Learning Disability
  Retardation
  Slow Learners
  Social/Emotional Maladjustment
  Thalidomide
  Trainable Mentally Retarded
Hawthorne Effect 88 (BR)-(FR), mid (MILL); 89 (HO)-(FELD); 93 (ALL); 232 mid (FELD); 233 (HO) to 235 (FR).
Heredity 290 bot (STAN).
Hypothesis, Objectives, and Expectations 53 bot (FELD); 54 (HA), (FELD); 103 mid (MILL) to 105 (BR); 160 (FR); 181 top (MA), (FR). See also Expectancy.

Individualization 57 mid (MA)-(PRO).
Inductive Validation 138 (FELD); 140 (HO).
Infant Education 204 (FR).
Inferential Statistics 31 bot (MA)-(ALL); 194 top (MA)-(ALL); 195 (BR); 196 (MILL)-(BR).
Informal Test Data 113 (MILL).
Instruction Evaluation 112 bot (MILL).
Instructional Materials 12 (FR)-(PRO); 13 (BR); 14 (PRO); 15 (FR); 37 (MILL); 76 (PRO): 197 bot (HO).
Instructional Materials Research and Development 235 top (FELD).
Integral Scaling 262 mid (STAN).

Intelligence Tests 210 above mid (FR).
Interaction 214 mid (HO), (FELD); 215 (MILL); 261 mid (MILL)-(STAN); 272 (BR); 273 top (STAN). See also Aptitude-Treatment Interaction.
Internal and External Validities 235 bot (FELD).
Interpretation 50 mid (BR); 133 (MILL).
IQ 262 mid (STAN); 294 mid (STAN).
IQ Tests 210 below mid (FR).
IQ Training 77 top (STAN); 78 (STAN).
Item Difficulty 81 (STAN).

Labeling 88 (BR) to 92 (FELD); 94 bot (HO) to 97 (ALL); 235 top (FELD).
Labels See Sampling.
Learning Disabilities 8 (HA); 55 top (PRO); 56 top (PRO); 96 (MA); 97 (MA); 166 (HA); 213 mid (MA).
Longitudinal Research 39 (FELD) to 40 (HA).

Maladjusted See Disturbed, Emotionally Disturbed.
Mastery Learning See Criterion-Referenced Measurement
Mastery Teaching 70 top (BR).
Matching 230 (BR)-(FELD), (MILL)-(FELD); 259 mid (STAN); 262 (MILL)-(STAN); 263 below mid (MILL); 264 top (MILL), mid (MILL); 265 top (STAN)-(MA); 266 above mid (STAN) to 267 (ALL); 283 (STAN); 284 (STAN); 285 (MILL); 292 above mid (MILL)-(STAN); 293 top (MILL)-(STAN), bot (MILL); 295 (MILL).
Matching and Control Groups 223 (ALL) to 225 (BR).
Materials Research and Development 13 bot (FELD); 14 (FR); 15 (FR); 46 (FELD); 188 (FELD) to 189 (ALL).
Measurement Errors 160 bot (STAN); 167 (MILL)-(STAN); 262 bot (STAN); 264 (HO), bot (STAN); 267 mid (MILL); 284 (STAN); 286 top (STAN); 291 mid (MILL); 292 bot (STAN); 293 bot (STAN); 294 top (STAN) to 95 top (STAN).
Measurement Error, Regression 295 mid (STAN).
Medical Results 59 (STAN); 60 mid (MA), below mid (FELD).
Mental Retardates 159 mid (FR); 259 top (MA)-(STAN)-(MA); 261 (HO); 263 top (MA)-(MILL)-(MA), bot (MA); 264 top, mid (MILL); 265 top (MILL); 266 above mid (STAN); 290 bot (STAN) to 291 mid (MILL); 195 (MILL).
Microteaching 123 bot (MILL).
Minority Groups 170 (MILL). See also Ethnic Groups.
Missing Data 270 mid (PRO); 273 (PRO)-(ALL); 287 mid (STAN); 296 bot (STAN).
Mongoloid 96 top (HA); 97 top (FELD); 260 below mid (STAN), bot (STAN); 263 top (MA); 289 bot (STAN); 290 bot (MA).
Multiple Correlation 245 bot (FELD), (FELD). See also Reliability.
Multiple Criteria 180 (BR); 217 (BR).
Multiple Regression 271 top, bot (STAN) to 272 (HO), (BR); 273 (HO); 277 (STAN).

Multivariate Statistics 201 bot (FELD).

National Assessment of Educational Progress (NAEP) 194 top (MA).
Negative Results 182 top (MILL)-(MA)-(MILL).
Nesting 260 below mid (STAN).
Neurological Impairment 34 (HO).
Normalize 278 below mid (STAN).
Norm-Referenced Measurement (NRM) 167 (STAN); 268 (CR); 169 top (FR), bot (HO_; 170 (MILL).
Norms 139 top (FR); 141 top (HO); 145 bot (HA) to 146 (ALL); 147 (BR) to 149 (ALL); 151 (HO) to 152 (MILL).
Nonisolated Variability (Error) 57 top (MA); 208 top (PRO)-(CR)-(PRO); 209 mid (MA).
Nonparametric Statistics 274 mid (STAN).
Nonstandardized ("Fugitive") Tests 143 (FR) to 144 (FELD); 147 (BR).

Observation 8 (BR); 44 (BR); 202 (FELD) to 204 (ALL).
Observational Rating Schemes 197 bot (FELD); 200 bot (FELD) to 201 top (FELD).
Observation Techniques 118 (MILL).
Obtained Scores 283 (STAN).
Objective Hierarchy 64 mid (BR); 65 (FR); 68 (STAN).
Objectives 81 (STAN); 121 top (MA); 159 mid (FR)-(STAN); 162 (HO) to 165 (BR); 166 (MILL); 168 (MA), (CR), (MILL); 169 below mid (HO). See also Bloom's Taxonomy, Hypothesis.
Operant Conditioning 58 (FELD); 59 (STAN); 60 below mid (STAN).
Organismic Variables 229 (HO).
Orthogonal Designs 271 bot (STAN); 272 (STAN).
Outlier 273 bot (STAN); 284 (STAN); 286 bot (STAN); 287 (HO). See also Rejection of Data, Data Deletion.

Paper-and-Pencil Tests 139 (MA) to 140 (HO).
Pass-Fail 170 mid (STAN).
Perceptual Impairment 31 top (HO).
Perceptual Training 20 top (HA) to 21 top (MA); 53 mid (HA); 197 bot (HO).
Performance Contracting 107 bot (HA) to 112 (FELD).
Performance Measures 152 (BR).
Permanent Psychological Records 171 (HA).
Phenylketonuria 213 (HA).
Physically Handicapped 159 mid (FR).
Physical Performance 77 mid (STAN).
Plan and Design Experiments 237 mid (STAN).
Pooling Results 181 mid (FR).
Population Variability 284 (STAN).
Power 209 bot (HO); 214 above mid (HO)-(MILL)-(HO); 216 top (HO); 237 bot (STAN); 284 (STAN); 260 below mid (STAN); 261 top, bot (STAN).
Practical Significance vs. Statistical Significance 53 bot (FELD); 211 (BR) to 212 (MILL).

Practice Effects 293 top (STAN).
Precision Teaching 57 bot (PRO); 58 (FELD)-(MILL); 59 mid (FELD); 60 bot (FELD) to 61 top, mid, bot (MILL).
Prediction 67 (STAN); 273 (PRO); 274 bot (MILL); 275 (STAN); 277 (STAN).
Prediction, Gain 259 below mid (STAN).
Prediction of Vocational Success 119 bot (MILL).
Prediction, Regression 295 mid (STAN).
Prediction Validation 83 bot (STAN); 132 top (FR); 133 (MILL); 135 (MILL); 151 (MA)-(FR)-(MILL)-(FELD); 153 (MILL); 170 (FR).
Preschool Programs 77 top (STAN).
Prescription 57 mid (MA); 76 (FELD), (STAN)-(PRO). See also ATI and Effective Treatment w/non ATI.
Prescription Validation 57 mid (MA)-(PRO); 58 (FR); 59 top (FELD), (STAN); 60 below mid (STAN), bot (FELD); 61 below mid (MILL) to 62 top (HA), (STAN); 64 top (MA)-(MILL), (FELD); 65 (MILL)-(HA); 66 (HO), (MILL); 67 (STAN); 68 (STAN); 69 bot (STAN) to 70 top, bot (BR), 71 (ALL).
Program Evaluation 13 (HO)-(MA)-(FELD); 37 (MILL); 57 mid (MA); 101 top (Feld); 103 mid (MILL) to 106 (FR); 194 (BR); 196 (FELD).
Project TALENT 277 (FELD).
Proportion 260 above mid (STAN); 271 bot (STAN); 272 (BR) to 273 top (STAN).

Randomization 194 (STAN) to 195 bot (STAN); 208 (FELD); 209 top (MA)-(FR)-(MA); 210 bot (PRO); 211 (HO); 215 bot (PRO); 229 (HO); 230 (MILL); 263 bot (MA); 265 bot (STAN); 283 (STAN), 284 (STAN); 291 top (STAN)-(MILL), (HO)-(STAN); 292 bot (STAN); 295 (MILL); 296 top, bot (STAN). See also Sampling.
Randomized Blocks Design 229 (HO); 230 bot (FELD).
Rate of Mastery 168 bot (STAN); 170 (FR).
Rating Scales 93 bot (MILL) to 94 top (HO).
Ratio Scale 261 bot (STAN).
Rat Psychology 59 (STAN); 60 below mid (STAN).
Raw Scores 260 below mid (STAN); 261 (HO).
Reading 160 mid (HO); 162 top (BR).
Reading Disability 21 (FELD); 27 (MA); 101 top (FELD).
Reading Instruction 53 top (HA); 101 top (FELD).
Regression to Mean 259 (ALL) to 260 mid (STAN); 262 (ALL); 264 mid (STAN), (HO); 266 (ALL) to 267 (ALL); 284 (STAN) to 286 (ALL); 288 (ALL) to 296 (ALL).
Reification of Constructs 81 (STAN).
Rejection of Data 220 bot (MA)-(HO). See also Data Deletion, Outlier.
Reliability 9 (HO); 15 (FR); 51 (FR); 52 (FR); 83 bot (STAN); 149 (MILL); 166 (CR); 167 (CR), (STAN); 214 (FELD); 241 (ALL) to 245 mid (MILL); 284 (CR) to 254 (ALL); 264 mid (STAN); 267 (STAN), mid (MILL); 284 (STAN); 293 mid (STAN); 294 top, mid, bot (STAN); 295 top (STAN).
Remediation 103 top (MA).

Replicability 180 (HA); 182 (FELD); 183 (ALL); 184 bot (HA); 185 top (MA).
Replication 195 bot (HO); 217 (HO); 219 mid (HO)-(PRO)-(HO).
Reported Measures 260 below mid (STAN), bot (STAN); 261 mid (STAN).
Reporting Pupil Progress 54 (BR); 174 top (PRO) to 175 bot (HA).
Research Design, Techniques, and Problems — See also:
  Administrative Problems
  Analysis of Covariance
  Analysis of Variance
  Aptitude-Treatment Interaction
  Bias
  Bias Contamination
  Blocking
  Case Study
  Case Study Approach
  Cause-Effect Relationship
  Change
  Comparative Studies
  Contamination
  Contamination and Bias
  Control Groups
  Controlled "One-Day" Studies
  Covariance
  Defining Problem
  Descriptive Research
  Ethics of Experimentation
  Expectancy
  Expectancy Effects
  Expectations
  Experimenter Bias
  Factor Analysis
  Gain
  Generalizability
  Hawthorne Effect
  Hypothesis, Objectives, and Expectations
  Internal and External Validation
  Longitudinal Research
  Matching
  Matching and Control Groups
  Multiple Correlation
  Multiple Criteria
  Multiple Regression
  Nesting
  Nonisolated Variability (Error)
  Observation
  Organismic Variables
  Orthogonal Designs
  Plan and Design Experiments

- Population Variability
- Prediction
- Prediction, Gain
- Prediction, Regression
- Program Evaluation
- Randomization
- Randomized Blocks Design
- Replication
- Research Interpretation
- Sampling
- Sampling Errors
- Selection of Variables
- Small Sample Size
- Studies, Developmental
- Studies, Manipulative

Research Interpretation 194 bot (MA)-(STAN); 196 (MILL); 235 bot (FELD). See also:

- Anonymity
- Contradictory Results
- Dangers
- Dissemination
- Errors of Inference
- Generalizability
- Negative Results
- Practical Significance vs. Statistical Significance
- Replicability
- Sampling Errors
- Significance Tests
- Social Impact
- Value Judgments on Soundness of Research

Residual Gain Scores 259 mid (STAN); 263 mid (MA); 283 bot (HO).

Retardation 34 (HA); 90 (HA)-(FR); 91 (MILL); 94 bot (HO); 97 (MA); 141 mid (FR); 146 top (MA)-(FR); 148 bot (FR); 202 bot (MA); 203 (FR).

Retention Tests 160 bot (STAN); 170 top (STAN).

Robustness 9 (HO); 210 bot (PRO).

Sampling 15 (MA)-(HA); 33 (MA)-(FR); 55 bot (PRO) to 57 (MILL); 87 (ALL); 194 (STAN); 213 top (MILL), (HA); 291 bot (STAN). See also:

- Attrition
- Bias
- Cell Frequencies
- Contamination
- Data Deletion
- Data Estimation
- Data Reduction
- Missing Data

  Outlier
  Program Variability
  Randomization
  Rejection of Data
  Small Ns
  Small Sample Size
  Substitution of Data
  Unique Cell Frequencies
  Variability
Sampling Errors 196 (MILL).
Scaling 261 (HO)-(STAN), bot (STAN); 262 mid (STAN).
Science 71 (STAN); 159 bot (MILL).
Scoring 22 bot (MILL).
Screening 34 (HA); 95 mid (HA).
Selection of Tests 149 (MILL).
Selection of Variables 237 mid (STAN).
Self-Concept Tests 164 top (MILL).
Self-Help Skills 169 below mid (HO); 170 top (STAN).
Sequence of Instruction 24 (FELD), (MILL) to 27 (ALL).
Significance Tests 59 top (FELD).
Slow Learners 259 top (MA).
Small Ns 276 (FELD) to 177 (STAN).
Small Sample Size 184 top (FELD); 201 top (FR) to 202 mid (FR); 210 bot (PRO); 212 (MILL) to (ALL); 213 bot (MILL)-(MA)-(HA).
Social/Emotional Maladjustment 88 bot (FELD); 130 top (PRO).
Socail Impact 12 (HO); 179 (PRO) to 180 (FR).
Social Studies 159 bot (STAN).
Statistical Assumptions 271 bot (STAN).
Statistical Significance 39 bot (MA).
Statistics 194 (ALL) to 197 (ALL). See also:
  Adjusted Data
  Analysis of Covariance
  Analysis of Variance
  Attentuation of Correlation
  Attrition
  Attrition - Biased
  Attrition - Natural
  Bayes
  Bayesian Statistics
  Cell Frequencies in a Factorial Design
  Change
  Computer
  Confidence Intervals
  Correlation
  Covariance (See also "Reliability")
  Data Coding Errors
  Data Deletion
  Data Estimation

    Data Pooling
    Data Reduction
    Degrees of Freedom
    Descriptive Statistics
    Difference Indices
    Errors of Inference
    Factor Analysis
    Gain
    General Linear Hypothesis
    Inferential Statistics
    Interactions
    Measurement Error Regression
    Missing Data
    Multiple Correlation
    Multiple Regression
    Multivariate
    Normalize
    Nonparametric Statistics
    Outlier
    Pooling
    Power
    Practical Significance vs. Statistical Significance
    Prediction
    Proportion
    Raw Scores
    Regression to Mean
    Rejection of Data
    Residual Gain Scores
    Robustness
    Small Ns
    Statistical Assumptions
    Substitution of Data
    Transforming
    Unequal Cell Frequencies
    Unweighted Means
Structure of Intellect 64 bot (FELD).
Studies, Developmental 229 (HO)-(MILL) to 230 (BR).
Studies, Manipulative 229 (MILL); 230 (BR).
Subjective Judgment Data 113 (BR).
Substitution of Data 219 top (FELD). See also Data Estimation.
Summative Evaluation 113 (STAN); 195 (FELD).
Survey Tests 159 bot (STAN).

Task Analysis 76 (STAN).
Teacher Effectiveness 8 mid, bot (MILL); 118 (MA) to 122 (MILLO.
Teacher-Made Tests 133 (MILL); 135 (BR); 159 bot (STAN); 169 mid (FR).
Teaching to Test 78 (HA); 170 (FR). See also Training under Applied Research.

- Team Teaching 236 top (STAN).
- Test Abuse 170 (MA); 171 (MA).
- Testing/Screening: educational, developmental histories 166 (HA).
- Testing Time 70 (MA)-(BR).
- Test Item Generation 162 (MILL); 168 (MILL).
- Test Rate 57 bot (PRO).
- Tests 65 (FR). See also:
  - Bender Gestalt 51 bot (MILL).
  - Bennett Mechanical Comprehension Test, Form C 81 (STAN).
  - Binet 58 (FR); 77 mid (STAN).
  - Cornell Test of Critical Thinking 142 bot (MILL).
  - Devereaux Behavior Rating Scale
  - Differential Aptitude Tests (D.A.T) 135 (FR), (MILL).
  - Eysench's Personality Inventory 79 bot (STAN).
  - Holland's Interest Inventory 79 bot (STAN).
  - Metropolitan Readiness Test 131 (MILL)-(HA); 134 (BR).
  - Minnesota Teacher Attitude Inventory 119 (MA) to 120 top (MILL).
  - Otis-Lennen 83 bot (STAN).
  - Peabody Individual Achievement Test (PIAT) 148 (HO)-(MA).
  - Pictorial Test of Intelligence 141 bot (FR).
  - Primary Mental Ability Test (PMA) 80 mid (BR); 81 (STAN); 83 bot (STAN).
  - Raven's Progressive Matrices 77 mid (STAN); 79 bot (STAN); 81 (STAN).
  - Rohrshach Technique 78 (HA).
  - Schematic College Aptitude Test (SCAT) 69 mid (STAN); 164 (STAN).
  - Sequential Test of Educational Progress (STEP) 71 (STAN).
  - Stanford Achievement Test (SAT) 170 mid (STAN).
  - Wechsler Bellevue Scale 145 (HO).
  - Wechsler Intelligence Scale for Children (WISC) 131 top (HA); 141 top (HO).
  - Wechsler Preschool Primary Scale of Intelligence (WPPSI) 139 (FELD).
  - Wepman Auditory Discrimination Test 131 top (HA).
- Tests and Testing — see also:
  - Achievement
  - Affective Tests
  - APA Standards
  - Aptitude Tests
  - "Cash Validity"
  - Ceiling Effects
  - Ceiling Tests
  - Change
  - Cognitive Tests
  - Computer
  - Constructs
  - Construct Validation
  - Content Validation
  - Creativity
  - Criterion-Referenced Measurement

Critical Reviews
Cross Validation
Cultural Bias
Cultural Deprivation
Diagnosis
Diagnostic Interviews
Diagnostic Tests
Difference Indices (CRM)
Domain-Referenced Measurements
Estimated True Score
Experimental Editions
Floor Disadvantage
Floor Effects
Fractionation of Assessment
Gain
Grading
Inductive Validation
Informal Test Data
Integral Scaling
Intelligence Testing
Internal and External Validities
Interpretation
IQ
IQ Tests
Item Difficulty
Mastery Learning
Mastery Teaching
Measurement Errors
Measurement Error Regression
Minority Groups (See Norms)
National Assessment of Educational Progress (NAEP)
Nonstandardized ("Fugitive") Tests
Norm-Referenced Measurement (NRM)
Norms
Observational Rating Schemes
Observation Techniques
Obtained Scores
Paper-and-Pencil Tests
Pass-Fail
Performance Measures
Physical Performance
Practice Effects
Prediction Validation
Rate of Mastery
Rating Scales
Ratio Scales
Raw Scores
Reliability

  Reported Measures
  Reporting Pupil Progress
  Results Interpretation
  Retention Tests
  Scaling
  Scoring
  Screening
  Selection of Tests
  Self-Concept Tests
  Structure of Intellect
  Subjective Judgment Data
  Survey Tests
  Teacher-Made Tests
  Teaching to Test (See also "Training" under Applied Research)
  Test Abuse
  Testing/Screening; educational, developmental histories
  Testing Time
  Test Item Generation
  Test Rate
  Tests - (See listing above)
  Test Time
  Transfer Measurements
  Types of Tests
  Unobtrusive Measures (Modes of Testing)
  Validity
  Variability
Thalidomide 25 mid (MA).
Theories of Learning 31 bot (HO); 38 bot (FELD).
Training Mentally Retarded 169 below mid (HO)-(MA); 170 top (STAN).
Training 49 mid (MA) to 53 below mid (HA); 109 bot (FR) to 110 top (FELD).
Training Abilities 80 top (BR).
Training Intelligence 107 bot (HA) to 108 mid (MA).
Transfer 160 mid (HO).
Transfer Measurements 110 (FELD).
Transforming 261 mid (STAN).
Types of Tests 247 (MILL) to 248 (FELD).

Unequal Cell Frequencies 170 (MA)-(PRO), (FELD)-(PRO); 272 (BR); 273
  top (STAN).
Unobtrusive Measures (Modes of Testing) 152 (MILL)-(HO)-(FELD).
Unweighted Means 170 mid (PRO).

Validity 14 (PRO); 51 mid (HA); 52 (FR); 130 top (PRO) to 139 top (FR);
  140 (HO); 141 (PRO) to 142 (ALL); 144 (PRO) to 145 (MA); 149 (MILL);
  247 top (BR)-(MA). See also:
  "Cash Validity"
  Construct Validation

    Content Validation
    Cross Validation
    Inductive Validation
    Internal and External Validation
    Prediction
Value Judgments on Soundness of Research 105 top (PRO) to (ALL).
Variability 167 (CR); 264 bot (STAN) to 265 top (MILL); 294 top (STAN)-(HO), bot (STAN). See also Sampling, Population Variability.